Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2016, **7,** 463–471 doi: 10.1111/2041-210X.12495

Optimizing sampling approaches along ecological gradients

Andreas H. Schweiger[1]*, Severin D. H. Irl[1], Manuel J. Steinbauer[1,2], Jürgen Dengler[3,4] and Carl Beierkuhnlein[1]

[1]*Biogeography, BayCEER, University of Bayreuth, Universitaetsstraße 30, 95440 Bayreuth, Germany;* [2]*Section Ecoinformatics & Biodiversity, Department of Bioscience, Aarhus University, Ny Munkegade 116, 8000 Aarhus, Denmark;* [3]*Plant Ecology, BayCEER, University of Bayreuth, Universitaetsstraße 30, 95440 Bayreuth, Germany; and* [4]*Synthesis Centre (sDiv), German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany*

Summary

1. Natural scientists and especially ecologists use manipulative experiments or field observations along gradients to differentiate patterns driven by processes from those caused by random noise. A well-conceived sampling design is essential for identifying, analysing and reporting underlying patterns in a statistically solid and reproducible manner, given the normal restrictions in labour, time and money. However, a technical guideline about an adequate sampling design to maximize prediction success under restricted resources is lacking. This study aims at developing such a solid and reproducible guideline for sampling along gradients in all fields of ecology and science in general.

2. We conducted simulations with artificial data for five common response types known in ecology, each represented by a simple function (no response, linear, exponential, symmetric unimodal and asymmetric unimodal). In the simulations, we accounted for different levels of random and systematic error, the two sources of noise in ecological data. We quantified prediction success for varying total sample size, number of locations sampled along a spatial/temporal gradient and number of replicates per sampled location.

3. The number of replicates becomes more important with increasing random error, whereas replicates become less relevant for a systematic error bigger than 20% of total variation. Thus, if high levels of systematic error are indicated or expected (e.g. in field studies with spatial autocorrelation, unaccountable additional environmental drivers or population clustering), continuous sampling with little to no replication is recommended. In contrast, sampling designs with replications are recommended in studies that can control for systematic errors. In a setting that is characteristic for ecological experiments and field studies strictly controlling for undeterminable systematic error (random error ≥10% and systematic error ≤10% of total variation), prediction success was best for an intermediate number of sampled locations along the gradient (10–15) and a low number of replicates per location (3).

4. Our findings from reproducible, statistical simulations will help design appropriate and efficient sampling approaches and avoid erroneous conclusions based on studies with flawed sampling design, which is currently one of the main targets of public criticism against science.

Key-words: ANOVA, curve fitting, ecological experiment, experimental design, model selection, regression analysis, replication, sampling design, simulation

Introduction

Temporal and spatial gradients are an intriguing and common feature in nature as already realized by the Greek philosopher Heraclites in 500 B.C. and expressed by Plato as '*panta rhei*' (everything flows) (Diels 1895). Major advances in the understanding of our current world have been made through analysing gradients. This is true for all disciplines of science but especially for ecology where, for decades, scientists have dealt with biotic responses along environmental gradients (Ramenskij 1918; Gleason 1939; Curtis & McIntosh 1951; Whittaker 1967; Palmer & White 1994; Sanders & Rahbek 2012).

Numerous concepts in ecology are based on continuous changes of biotic features along abiotic environmental gradients such as the niche concept (Grinnell 1917), coenoclines (Gauch & Whittaker 1972), the intermediate disturbance hypothesis (Connell 1978) or the stress-gradient hypothesis (Bertness & Callaway 1994). Efforts were made in the last decades to develop and test analytical techniques for characterizing single species and whole community responses along environmental gradients in a sound and reproducible manner.

*Correspondence author: E-mail: andreas.schweiger@uni-bayreuth.de

These analytical approaches include the establishment of similarity indices (e.g. Bray & Curtis 1957), univariate regression and multivariate ordination methods (Whittaker 1967; ter Braak & Prentice 1988) as well as the methodological concepts of beta-diversity (Whittaker 1972) and species response curves (Austin 1987; Huisman, Olff & Fresco 1993). Numerous studies exist on how to distribute samples through space and time to optimal cover the underlying ecosystem's variability (e.g. Gillison 1984; Legendre *et al.* 1989; Stein & Ettema 2003). Important methodological concepts like the response surface methodology (Box & Wilson 1951) evolved from the need to increase the prediction success of gradient patterns in natural science (Myers, Khuri & Carter 1989).

Based on the gathered knowledge from all this original research, numerous text books give recommendations about how to plan and conduct ecological sampling and how to analyse the sampled data in order to approximate the underlying pattern as close as possible (e.g. Cochran & Cox 1957; Gregoire & Valentine 2007; Lohr 2009; Gotelli & Ellison 2013). However, a clear guideline about how many samples are needed in which intensity along a gradient under study to most efficiently and accurately identify ecological patterns along the studied gradient is still missing.

Several authors already provide technical assistance to improve sampling with the aim to increase the reliability of results obtained from the sampled data. Adequate total sample size can be estimated using pre-studies (Eckblad 1991) and/or *a priori* power tests (Bartlett, Kotrlik & Higgins 2001; Ioannidis 2005; Bakker, van Dijk & Wicherts 2012). By contrast, for estimating the number of necessary replicates, only approximations based on empirical observations, such as the 'rule of ten' by Gotelli & Ellison (2013), are available. This rule of thumb suggests a minimum number of ten observations per sampling point. However, the authors themselves note that '[...] *the rule of ten is not based on any theoretical principle of experimental design or statistical analysis, but is a reflection of our hard-won field experience with designs that have been successful and those that have not*'. Even less is known when it comes to balancing the number of replications per point of observation against the number of observation points along a spatial or temporal gradient of interest although this is a major source of error in designing ecological studies (Hurlbert 1984; Quinn & Keough 2002). Based on this lack of information, there are calls for a clear, empirically based guideline about how to optimize ecological sampling (e.g. Bartlett, Kotrlik & Higgins 2001) in order to conduct cost-efficient but still statistically sound analyses. This is especially important as sampling is cost-, time- and/or labour-intensive and, thus, strongly restricted by limited funding which is characteristic for almost every scientific study.

The two main characteristics in sampling design are total sample size and the number of replicates per sampling point along the gradient under study (Gotelli & Ellison 2013). The necessity to take an adequate total number of samples results from the fact that the reliability of findings depends on the total sample size in relation to the random variation that can mask the focal pattern (Eckblad 1991; Bartlett, Kotrlik & Higgins 2001). The replication of observations at each sampled location along a spatial/temporal gradient of a certain environmental factor (e.g. spatial or temporal variation of temperature, pH; hereafter called predictor level) follows two aims: (i) to increase the accuracy of parameter estimation and (ii) to provide information on the natural variation within the data set on which the statistical tests for differences between the predictor levels are applied (Southwood & Henderson 2000; Quinn & Keough 2002). It is obvious that the higher the total sample size and the higher the number of observations per predictor level (*n*; replicates; for sake of linguistic simplicity, we use replicates for *n* despite sometimes in the literature it is also applied to $n - 1$), the more precisely one can estimate the underlying pattern. If the number of total observations is constant, there is an inevitable trade-off between the number of observations which can be sampled per predictor level and number of sampled predictor levels along the gradient of interest. Up to now there has been no technical guidance about how to balance the number of predictor levels and the number of replicates when aiming for maximum prediction success under limited resource (i.e. total sample size).

For the study of a response variable along a gradient of a certain predictor, practically any solution ranging from only two predictor levels with many replicates to many predictor levels with no replication can be found in the recent literature (Scheiner & Gurevitch 2001; Quinn & Keough 2002; Gotelli & Ellison 2013). Ecologists are usually interested in differences of biotic response under certain environmental settings (traditionally experimental ecologists) or study the actual shape of a biotic response along the gradient of a certain environmental factor (field- and macro-ecologists). Based on these two different ways of studying ecological response to environmental changes, two major methodological approaches are common in current ecological research. Field- and macro-ecologists tend to sample gradients continuously (in a systematic or preferential manner) but without replication ('regression approach': Mac Nally 2000; Quinn & Keough 2002). In contrast, experimental ecologists traditionally use replicated sampling of two to few predefined predictor levels ('ANOVA approach': Cottingham, Lennon & Brown 2005; Beier *et al.* 2012). However, also in experimental ecology, recently a call for 'regression-based experimental design' has been launched that comes along with reduced replicates but higher numbers of predictor levels (Cottingham, Lennon & Brown 2005; Beier *et al.* 2012; de Boeck *et al.* 2015). However, no feasible methodological recommendation exists for this way of conducting ecological experiments so far.

Under natural, non-experimentally controlled conditions biological systems are characterized by high random variation, which will likely dilute the underlying relationships of interest (Quinn & Keough 2002; Lohr 2009). Furthermore, data from field investigations can be affected by a complex interplay of various interacting or opposing gradients. Such factors can be seen as systematic errors in the biological response along a gradient under study and can hamper the study of responses to one specific environmental gradient (Gauch & Whittaker 1972; Richardson *et al.* 2012; Steinbauer *et al.* 2012). Thus, measurements in natural systems are always subject to errors and uncertainties related to measurement errors, ecological and environmental stochasticity and unaccounted, additional influencing factors (Taylor1991; Clark 2003; Richardson *et al.* 2012).

Using artificial data instead of ‘real-world’ data allows excluding or adding known random variation and systematic errors to the ‘observational’ data. In this paper we use simulations based on artificial data with known properties in terms of random and systematic noise to address the problem of how to balance the number of predictor levels sampled along a gradient of a certain environmental factor and number of replicates per predictor level in order to maximize prediction success of the underlying ‘true’ pattern. By varying random as well as systematic noise in the data, we provide a statistically sound and reproducible guideline about how to optimally sample ecological data, which is applicable in all fields of ecology and science in general.

In our simulations, we assume that a ‘true’ relationship between gradual changes of an environmental factor and the ecological responses thereon follows a defined response shape that corresponds to a common mathematical relationship. We add random and/or systematic errors of different degrees to the data set, which can mask the ‘true’ relationship. Then, we draw samples from the simulated gradients by using different sampling approaches and compare the results to the underlying, ‘true’ pattern.

Materials and methods

MODEL SELECTION AND ARTIFICIAL DATA CONSTRUCTION

We simulated five response shapes frequently occurring in ecological studies (no response as a null model/control, linear response, exponential decay, unimodal response with centred maximum and unimodal response with non-centred maximum) based on simple linear models (see Table 1; inlets in Appendix S1). The response variable (y) thereby represents any kind of biotic response (e.g. species richness, photosynthetic rate, biomass, phylogenetic diversity) that varies along a gradient of the predictor variable (x), which, in turn, represents any kind of spatial or temporal change of environmental conditions (e.g. spatial/temporal change of temperature, pH, disturbance intensity). To make our inferences easily transferable to any response and predictor variable independent of the studied system, we scaled our parameters in arbitrary units (predictor variable from 0 to 1000, response variable from 0 to 1).

Table 1. Response shapes considered in the comparison and functions used for their implementation as well as parameterization in the simulations

Response shape	Function	a	b	c	# of parameters
No response	$y = c$			0·5	1
Linear	$y = ax + c$	−0·001		1	2
Exponential decay	$y = \exp(-x^{a})$	0·3			1
Unimodal centred	$y = ax^2 + bx$	$-4{\cdot}10^{-6}$	$4{\cdot}10^{-3}$		2
Unimodal non-centred	$y = ax^2 + c$	$-1{\cdot}10^{-6}$		1	2

Data sampled from natural systems always include errors and uncertainty (Taylor 1991). Traditionally, two types of error can be classified depending on their way of affecting the sampled data and, thus, the statistical inference drawn from it (Richardson *et al.* 2012). Whereas the so-called random error combines non-directional noise which influences the response variable in addition to the main predictor in a stochastic and, thus, unpredictable way (observed value = expected value + random noise; Fig. 1a), ‘systematic error’ summarizes a bias in the data which is constant but unknown (Abernethy, Benedict & Dowdell 1985). This systematic error may originate, for example from spatially clustered environmental characteristics, population effects or any other non-accounted or non-accountable influential factor (observed

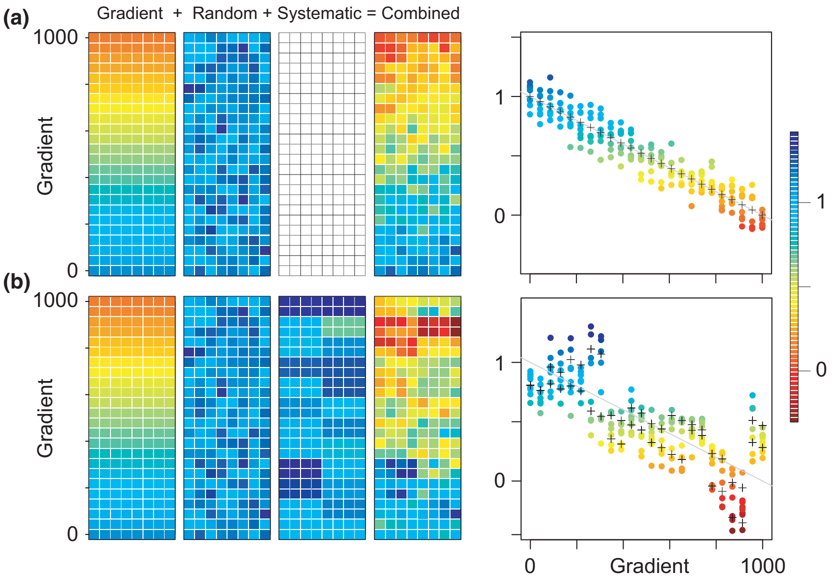

Fig. 1. Artificial data for (a) only random error or (b) random and systematic error. The observed value was combined from the sum of (i) an expected value from a gradient, (ii) a random error (here 10% of gradient length) and (iii) a systematic error (here 20% of gradient length). The eight grid cells per considered predictor level along the gradient (l = 24) represent eight samples (n = 8) considered for each predictor level. In the right panel, the grey trend line indicates the gradient (expected value without errors) and ‘+’ expected values without random error (gradient and systematic error). The points represent the samples drawn at each predictor level along the gradient (equal grid cells of ‘Combined’).

value = expected value + random noise + systematic influence; Fig. 1b).

In order to reflect different levels of random variation in the simulated data, we assumed that the observed values of the response variable are scattered around the expectancy value of a certain predictor level with a normal distribution that corresponds to a standard deviation (*sd*) of 0·02, 0·05, 0·1, 0·15, 0·2 and 0·25 units, that is 2% to 25% of the total variation. Information about the levels of random noise in 'real-world' data is extremely rare, and only very few studies explicitly focus on the quantification of random noise in ecological data. Based on sampling designs to explicitly quantify random noise in eddy flux measurements, a highly uncertain method in environmental science, Richardson *et al.* (2012) estimated random noise to reach a maximum of 23% of total variation. Similar levels of random noise were quantified by Kelly *et al.* (2009) for an ecological quality index for rivers based on the community composition of diatoms where random noise varied between 3 and 22% of total variation (on average 11·3 ± 4·6%). As the levels of random noise observed in both studies are completely covered in our simulations, we believe that our simulations will be of practical use in many 'real-world' situations.

In addition to the non-directional effect of the random error, we added different levels of systematic error to account for factors, which are not yet covered in the actual study but have a directional effect on the observed response pattern. This systematic bias added to the data can be considered as a 'fully systematic error' (c.f. Richardson *et al.* 2012) as it influences all drawn samples to the same degree. To implement this additional, structured noise in our data, we randomly shifted the 'true' expectancy value y_i independently at each predictor level x_i sampled along the gradient by 0% to 25% of total variation (*sd.err* = 0, 0·02, 0·05, 0·1, 0·15, 0·2 and 0·25 units, respectively). Thus, the extent by which a certain level of systematic error shifted the 'true' expectancy values y_i at a certain predictor level x_i was similar for all simulations but differed in the direction (whether y_i was over- or underestimated) for the different simulation settings (e.g. different response shapes, levels of random variation). Afterwards, we added different levels of random error in addition to the systematic error by sampling normally distributed around the new, shifted expectancy values (again random variation *sd* of 0·02, 0·05, 0·1, 0·15, 0·2 and 0·25 units, respectively, Fig. 1b).

We considered 31 different values for *total number of observations* (*N*) with a minimum sample size of 6 followed by a stepwise increase of total samples size from 10 to 300 in steps of 10 (i.e. $N = 6, 10, 20, \ldots, 300$). In combination with the six different levels of random and seven levels of systematic errors, respectively, we tested a total of 1302 combinations of study settings. In agreement with common recommendations in ecological literature about gradient analysis in ecology (e.g. Kenkel, Juhász-Nagy & Podani 1989; Quinn & Keough 2002), we placed the sampled predictor levels evenly (equidistantly) along the gradient of 0–1000, with the two end points always being sampled. The *number of predictor levels* (*l*) ranged from 2 to the total number of observations (*N*). The *number of observations per predictor level* (*n*) varied from one observation (no replication) per predictor level ($n = 1$) to 50% of total number of observations ($n = N/2$). For each value of *N*, all whole-number factorizations $N = n \cdot l$ were considered. For example, if the total number of observations was $N = 6$, we compared three different sampling strategies: 6 predictor levels with 1 observation per level, 3 predictor levels with 2 replicates per level and 2 predictor levels with 3 replicates per level.

EFFECT OF SAMPLING APPROACH ON CORRECT PATTERN IDENTIFICATION

The data set for each parameter combination (response shape × level of random variation × level of systematic error × total number of observations × number of observations per predictor level) was subjected to a simple linear, one-factorial regression analysis between a response variable *y* (e.g. species richness) and a predictor variable *x* (e.g. spatial/temporal change of temperature) with the five response shapes (transformations of the predictor x) of Table 1 to choose from by using the *lm*() command in R (v. 3.0.1, R Development Core Team 2013). Replications sampled for the particular predictor levels were thereby treated as independent observations. The most appropriate model among the five options was then selected based on AICc, which takes model complexity and total sample size into account (Burnham & Anderson 2002). For each parameter combination, we repeated sampling and the subsequent analyses 1000 times.

In a next step, the statistically inferred response shape was compared to the 'true' response shape in two ways. In the *correct model* approach, we calculated the fraction of correctly detected response shapes (irrespective of the model parameters). We therefore defined a pattern to be correctly predicted, when the response shape chosen from the algorithm based on AICc was the same as the predefined, 'true' response shape. We calculated the fraction of correctly predicted response shapes for each combination of *N*, *l* (*n*), *sd* and *sd.err* from 1000 model runs.

The *precision of prediction* approach quantified how much the inferred response shape deviated from the actual response shape (irrespective of the function type). For this second approach, we calculated the absolute deviation of the predicted response shape from the 'true' response shape by using a numerical integration approach. Therefore, we summed up the mean absolute differences between the predicted and true response value ($|\hat{y}_i - y_i|$) for a defined number of predictor levels along the gradient under study with $x_i = 0, 10, 20, \ldots, 1000$ and divided this sum by the number of considered predictor values (101). As a result, the area enclosed between the 'true' and the predicted response shape along the whole gradient under study standardized by the number of sampled locations along this gradient is calculated. The derived values were then divided by the maximum deviation between inferred and true response pattern, which could be observed among all five response shapes for the respective level of random variation. The complement of these standardized values (i.e. 1 – value) increases with increasing precision towards 1 and was defined as *precision of prediction* (POP).

Besides the type II error (chance of failing to detect a present pattern) which is captured by these first two approaches, we also captured the type I error problematic (chance of detecting a non-existent pattern) in a third approach. Using the no-response pattern as a basis, we calculated the fraction of cases where patterns were erroneously detected from 1000 model runs for each combination of *N*, *l* (*n*), *sd* and *sd.err*.

To visualize the simulation results, we plotted the fraction of correctly detected response shapes and the POP values for each of the five response shapes as well as a mean of these (excluding the no-response pattern) as a function of the total number of observations, the number of predictor levels and the number of replicates. Trend surfaces for the visualizations were fitted by using least-squares based on a third-order polynomial (*surf.ls*() and *trmat*() commands of the spatial package in R, v. 7.3-7; Venables & Ripley 2002) as well as isolines (*contour*() command implemented in R). The same was done for the fraction of erroneously detected pattern detection based on the no-response pattern.

All simulations and calculations were conducted in R with the add-on packages CATOOLS (v. 1.14, Tuszynski 2012) and AICCMODAVG

(v. 1.35, Mazerolle 2013). In order to handle the large computational capacity required to calculate the simulations, the MULTICORE package (v. 0.1-7, Urbanek 2011) was used to run parallel computations of simulations on a multiple core server with a Linux operating system. Statistical relationships were tested with Pearson correlation analyses as well as simple linear models with a level of significance of alpha = 0·05. Visualization was supported by the R packages FIELDS (v. 7.1., Nychka, Furrer & Sain 2014) and CLASSINT (v. 0.1-21., Bivand 2013). The scripts we implemented in R for simulation can be found in the electronic appendix.

Results

The prediction success, expressed as the probability to detect the correct response shape (*correct model* approach) and the precision of this prediction (POP), was strongly related to the number of predictor levels *l* and the number of replicates per predictor level *n* (see Fig. 2 as well as Appendix S1). We observed a strong increase in the prediction success from 2 to about 10 predictor levels, almost independent of the total number of observations and the level of random variation. Except for a very low total number of observations below N = 10 and the no-response pattern, this was true for all levels of random variation (*sd*) and all tested response shapes in both approaches (Appendix S1 and S2).

The correct model was detected in more than 90% of cases for all parameter combinations and response shapes within a range of *l* from about 10–30 (for exact values, see Appendix S2a, c, e and S3). We observed this optimal range of number of predictor levels for all levels of random variation (*sd*). The same range of *l* was also true for the *precision of prediction* approach, in which the POP values reach the maximum of 1 (Appendix S2b, d, f). With further increase of *l*, concomitant with decreasing *n*, however, the probability of correct prediction decreased again. Thus, an intermediate number of predictor levels of *l* = 10–15 turned out to be optimal, regardless of the total number of observations and level of random variation along the sampled gradient. This optimal, intermediate number of predictor levels holds true for low to intermediate levels of systematic errors below 15% of total variation. For a systematic error ≥15%, the adverse effects of a higher number of predictor levels disappeared. For high levels of systematic error, prediction success reached its maximum at *l* ≥ 10 and stayed constant with increasing *l*, although overall prediction success was lower for higher levels than for lower levels of systematic error (Fig. 2).

Besides the number of predictor levels, also the number of replicates per predictor level *n* strongly affected the prediction success. Although the overall prediction success decreased with increasing levels of random variation, the need to take replicates in order to maximize prediction success significantly increased with increasing levels of random variation. While

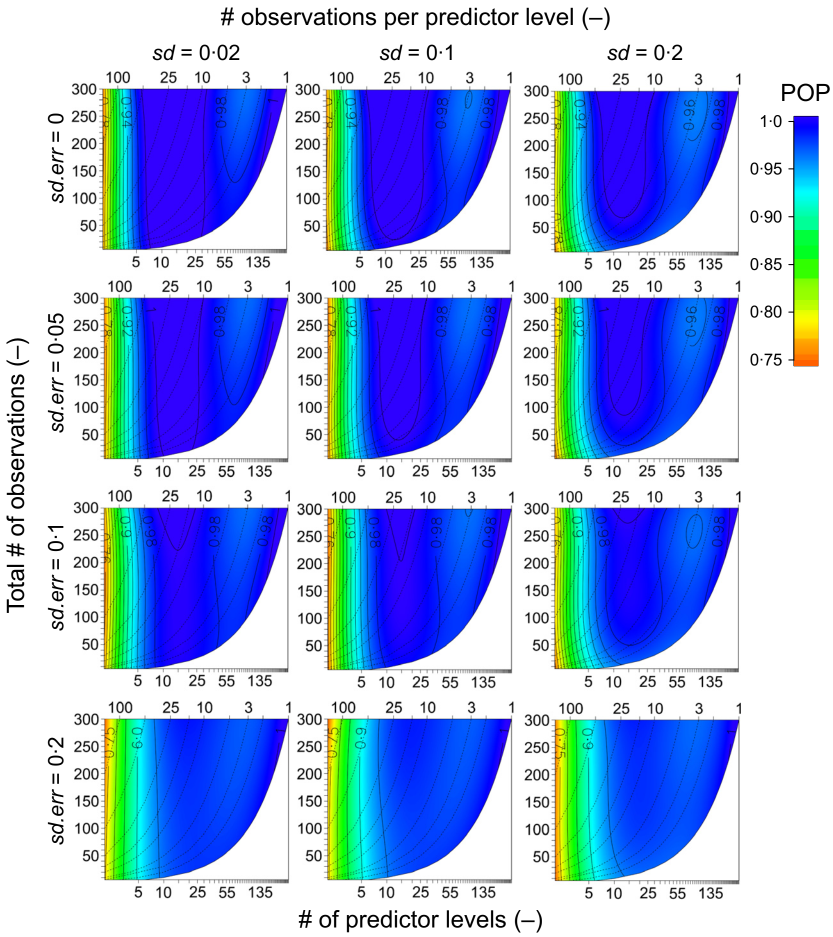

Fig. 2. Prediction success depicted as the *precision of prediction* (POP) for three different levels of random variation (*sd* = 2%, 10% and 20% of total variation) and four different levels of systematic error (*sd.err* = 0%, 5%, 10% and 20% of total variation) based on an average of the tested response patterns (without no-response). POP values dependent on the total number of observations, number of predictor levels and number of observations per predictor level are depicted. Solid lines show isolines for a selection of POP values, and dashed lines show isolines for a selection of number of observations per predictor level.

no replication was needed to reach the highest possible prediction success for low levels of random variation (*sd* = 0·02 and 0·05), it was indispensable at higher levels of random variation (*sd* ≥ 0·1, see Appendix S1). Our simulations show that on average *n* = 3 observations per predictor level in combination with *l* = 10–15 predictor levels along the gradient are appropriate to achieve the highest possible prediction success for a systematic error of ≤10% of total variation (see Fig. 2 for *sd.err* = 0–0·1 and *sd* = 0·02, 0·1 and 0·2). For these low to intermediate levels of systematic error, our simulations highlight an increasing need of replicates with increasing random variation to achieve the highest possible prediction success (i.e. POP = 1) with minimum effort (lowest possible *N*). This positive correlation between the level of random variation and the number of necessary replicates was significant under the absence of a systematic error (Pearson's *r* = 0·98) as well as for a systematic error of 2% (*r* = 0·99) and 5% (*r* = 0·98) of total variation (*P* < 0·001 and d.f. = 4 in all cases). The positive effect of increasing random variation on the number of necessary replicates increased with increasing systematic error up to 5% of total variation. However, replication was negligible when the systematic error exceeded 10% of total variation.

Regarding type I error problematic (chance of detecting a non-existent pattern), the probability to erroneously detect a pattern strongly increased with increasing systematic error (Appendix S4). Maximum probability for erroneous pattern detection was 30–32% when no systematic error was added to the data but strongly increased for 5% of systematic error. While increasing random error enhanced the probability of erroneous pattern detection under the absence of a systematic error (Appendix S4 for *sd.err* = 0), increasing random error mitigated the effect of an increasing systematic error on erroneous detection probability (Appendix S4 for *sd.err* = 0·05, 0·1 and 0·2). Under the absence of a systematic error (*sd.err* = 0), at least *n* = 2–3 replicates per predictor level were necessary to avoid an erroneous detection of inexistent patterns. This positive effect of taking replicates increased with increasing random error but strongly decreased with increasing systematic error (*sd.err* ≥ 0·05). Thus, a low to intermediate number of replicates decreased the risk of detecting inexistent patterns but intermediate to high levels of systematic error (≥5% of total variation) diminished this positive effect of replicates and significantly increased the risk of an erroneous detection of non-existent patterns, that is led to an inflation of the type I error.

Discussion

PREDICTION SUCCESS UNDER EXPERIMENTAL SETTINGS VS. FIELD CONDITIONS

Our simulations show that low numbers of predictor levels (points in space/time sampled along spatial/temporal gradients) in combination with medium to high numbers of replicates, that is the typical approach among experimental ecologists, may not be the most effective way to detect response shapes in environmental factors that continuously change along gradients. Our results suggest aiming at increasing the number of predictor levels and, in exchange, reducing but not abandoning replication if total sample size is restricted. In other words, it seems to be more advantageous for experimental ecologist studying gradients to move away from the approaches typically using two to three predictor levels and many replicates.

We recommend a similar sampling approach for field studies, which effectively control for high levels of systematic error by *a priori* excluding or at least minimizing additional interfering variables that might alter the underlying pattern of interest. Sampling a high number of predictor levels along the gradient under investigation with no replication at individual predictor levels, which is widespread among field- and macro-ecologists, may not always be an appropriate solution. Instead an intermediate number of predictor levels in combination with a low number of replicates (10–15 predictor levels and 3 replicates) seem to be in many cases a better road to prediction success when high levels of random variation (in our case ≥5% of total variation) and/or low to intermediate levels of systematic error (≤5% of total variation) can be expected. If, however, higher levels of systematic error are likely (in our case ≥10% of total variation), continuous sampling without replication becomes preferable compared to sampling fewer predictor levels along the gradient with replications. This is especially the case for field studies along gradients that do not explicitly control for additional disturbing factors such as biotic interactions (e.g. competition) which might alter the effect of the underlying abiotic driver of a biotic response (e.g. species occurrence or abundance).

Comparing a set of functions that corresponds to response shapes frequently found in ecology, we showed that an intermediate number of 10–15 predictor levels along the gradient under investigation in combination with three observations per predictor level maximize prediction success for intermediate to high levels of random variation (≥5% of total variation) and small to intermediate levels of systematic error (≤5%). Thus, taking ten replicates per predictor level, as recommended by Gotelli & Ellison (2013), based on field experience will likely cause unnecessary oversampling. This holds true for all response shapes tested in our study: the linear response, the exponential decay and the two unimodal response patterns with centred and shifted maximum.

IMPLICATION FOR FUTURE STUDIES IN ECOLOGY

The preservation of quality in scientific studies is of particular importance as 'unreliable research' is currently one of the main targets of public criticism against science (e.g. The Economist 2013). Several studies conducted in cancer research, neuroscience and psychology, which had high impact on society and economics, recently turned out to lack the required reproducibility (e.g. Prinz, Schlange & Arrowsmith 2011; Simmons, Nelson & Simonsohn 2011; Begley & Ellis 2012; Shanks *et al.* 2013; Open Science Col-

Box 1. Optimal gradient sampling in a nutshell for situations with one major gradient (or one factor to be tested) and when more complex response shapes than those of Table 1 are not expected.

1. In controlled environments (i.e. experiment-like settings): Intermediate number of points in space/time sampled along spatial/temporal gradients (10–15) and a low number of replicates per point (3) suggest a total sample size of 30–45. This approach is also sensible for field ecologists, if confidence is high that a possible systematic error (i.e. unknown additional predictor variables) is controllable.
2. Under field conditions (i.e. high levels of systematic error): If systematic errors are unaccountable or are likely to be high, gradual sampling with no replication should be preferred. However, predictor levels and sample size necessary to obtain high prediction success strongly increase with increasing systematic error (in our case on average >200).
3. Type I error inflation by systematic errors: The probability of an erroneous detection of a non-existent pattern (type I error) significantly increases with increasing systematic error (bias in the data which is constant but unknown).

laboration 2015), a major pillar of science. In most cases, the reason for this was low statistical power of the studies caused by small sample size and/or a small number of replicates (Bakker, van Dijk & Wicherts 2012; Begley & Ellis 2012; Button *et al.* 2013). However, flawed sampling design is not an exception, but seems to be a relatively widespread phenomenon in science (Ioannidis 2005). To counteract this problem, several authors strongly recommend a careful sampling design based on knowledge from previous studies (Legendre *et al.* 1989; Bakker, van Dijk & Wicherts 2012). Ioannidis (2005) suggests a general increase in sampling effort but hints at the same time at the associated rising costs. Our simulations offer a solid basis to further improve experimental and sampling design in ecological studies and, thus, may play an important part in contributing to save funds and labour without an associated loss of quality.

LIMITATIONS OF OUR SIMULATION APPROACH AND OUTLOOK

To maintain the straightforward message of our study, we had to restrict the tested simulation settings to relatively simple functions with only one major gradient (predictor). We assume that this covers the situation in a significant fraction of ecological studies as well as studies from other disciplines. However, the simulation and testing framework presented here could possibly be extended to more complex function types (like breakpoint or sigmoid functions; Matthews *et al.* 2014 or power functions; e.g. Dengler 2009) or to more than one predictor of interest (multifactorial or mixed effect models). The latter is particularly relevant in experimental studies where two or more factors are crossed (factorial design), but also in observational studies where often more than one environmental driver of biotic patterns interact.

Furthermore, we equidistantly placed the predictor levels (sampling locations) along the gradient under study. We, thus, did not consider the effect of preferential sampling along environmental gradients which is, for example applied by the gradient-oriented sampling (gradsect method; Gillison 1984) or the adaptive-sampling approach (Thompson & Seber 1996). As organisms are often not randomly distributed along environmental gradients but lump in preferential ranges (Fortin, Drapeau & Legendre 1989; Legendre *et al.* 1989), future studies should also elaborate on the effect of preferential sampling to further increase sampling efficiency. However, the need of detailed *a priori* knowledge about the ecological niche characteristics that is still lacking for many organisms may hamper preferential sampling. While all these mentioned topics can be seen as a limitation of the present study, our results provide a first, clear and reproducible guideline about how to optimize sampling along ecological gradients.

Our study did explicitly not implement standard goodness-of-fit approaches (like R^2, or f-statistic based p-values) as these measures are susceptible to systematic errors. Sampled data modified by systematic error might be perfectly predicted by a model, which does not match the 'true' underlying pattern. This is particularly true in scenarios with few predictor levels and many replicates where an erroneous model could still score high R^2 values, leading to entirely wrong conclusions. These erroneous conclusions caused on high values of standard goodness-of-fit approaches stress the importance of approaching scientific questions with consistent theory and quantifying possible sources for systematic errors.

Our simulations show that replication is inevitable in experimental studies and advisable for observational field studies, unless unaccounted systematic errors occur, potentially distorting the underlying pattern. This is especially true, when the random error is high. However, optimal sampling strategies have to be selected context-dependent and differ with the required accuracy, which has to be achieved, as well as the number (uni- or multivariate) and the mathematical character (discrete or continuous variables) of the variables tested in the particular study (Kenkel, Juhász–Nagy & Podani 1989). Although this context dependency of sampling strategies seems to impede general statements about optimal sampling strategies, a systematic and reproducible approach like ours could help to set clear framework conditions on which future studies could build on in order to further optimize sampling in ecology and possibly also in other scientific disciplines. Based on our results, we infer basic guidelines for gradient sampling in Box 1.

Acknowledgements

We would like to express our gratitude to the organizers of the GfÖ (Ecological Society of Germany, Austria and Switzerland) conference 2013 club night in Potsdam, Germany, for providing the authors with the necessary inspiration (and a sufficient amount of cereal-based beverages) for this study.

Data accessibility

This study was conducted with artificial data which can be reproduced by following the methods section of this manuscript or by using the R-script which can be found in the electronic appendix.

References

Abernethy, R.B., Benedict, R.P. & Dowdell, R.B. (1985) ASME measurement uncertainty. *Journal of Fluids Engineering*, **107**, 161–164.

Austin, M.P. (1987) Models for the analysis of species' response to environmental gradients. *Vegetatio*, **69**, 35–45.

Bakker, M., van Dijk, A. & Wicherts, J.M. (2012) The rules of the game called psychological science. *Perspectives on Psychological Science*, **7**, 543–554.

Bartlett, J.E. II, Kotrlik, J.W. & Higgins, C.C. (2001) Organizational research: determining appropriate sample size in survey research. *Information Technology, Learning, and Performance Journal*, **19**, 43–50.

Begley, C.G. & Ellis, L.M. (2012) Drug development: raise standards for preclinical cancer research. *Nature*, **483**, 531–533.

Beier, C., Beierkuhnlein, C., Wohlgemuth, T., Penuelas, J., Emmett, B., Körner, C. *et al.* (2012) Precipitation manipulation experiments – challenges and recommendations for the future. *Ecology Letters*, **15**, 899–911.

Bertness, M.D. & Callaway, R. (1994) Positive interactions in communities. *Trends in Ecology and Evolution*, **9**, 191–193.

Bivand, R. (2013) *classInt: choose univariate class intervals. R package version 0.1-21.* URL http://CRAN.R-project.org/package = classInt [accessed 11 August 2014]

de Boeck, H., Vicca, S., Roy, J., Nijs, I., Milcu, A., Kreyling, J. *et al.* (2015) Global change experiments: challenges and opportunities. *BioScience*, **65**, 922–931, doi:10.1093/biosci/biv099.

Box, G.E.P. & Wilson, K.B. (1951) On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society: Series B*, **13**, 1–45.

ter Braak, C.J.F. & Prentice, I.C. (1988) A theory of gradient analysis. *Advances in Ecological Research*, **18**, 271–313.

Bray, J.R. & Curtis, J.T. (1957) An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, **27**, 325–349.

Burnham, K.P. & Anderson, D.R. (2002) *Model Selection and Multimodel Inference – A Practical Information-theoretic Approach*, 2nd edn. Springer, New York City, New York, USA.

Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J. & Munafò, R. (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, **14**, 365–376.

Clark, J.S. (2003) Uncertainty in ecological inference and forecasting. *Ecology*, **84**, 1349–1350.

Cochran, W.G. & Cox, G.M. (1957) *Experimental Designs*, 2nd edn. Wiley, New York City, New York, USA.

Connell, J.H. (1978) Diversity in tropical rain forests and coral reefs. *Science*, **199**, 1302–1310.

Cottingham, K.L., Lennon, J.T. & Brown, B.L. (2005) Knowing when to draw the line: designing more informative ecological experiments. *Frontiers in Ecology and the Environment*, **3**, 145–152.

Curtis, J.T. & McIntosh, R.P. (1951) An upland forest continuum in the prairie-forest border region of Wisconsin. *Ecology*, **32**, 476–496.

Dengler, J. (2009) Which function describes the species-area relationship best? – a review and empirical evaluation. *Journal of Biogeography*, **36**, 728–744.

Diels, H. (1895) *Simplicius, In Aristotelis Physicorum Libros Quattuor Posteriores Commentaria.* Reimer, Berlin.

Eckblad, J.W. (1991) How many samples should be taken? *BioScience*, **41**, 346–348.

Fortin, M.-J., Drapeau, P. & Legendre, P. (1989) Spatial autocorrelation and sampling design. *Vegetatio*, **83**, 209–222.

Gauch, H.G. Jr & Whittaker, R.H. (1972) Coenocline simulation. *Ecology*, **53**, 446–451.

Gillison, A.N. (1984) Gradient oriented sampling for resource surveys – the gradsect method. *Survey Methods for Nature Conservation* (eds K.R. Myers, C.R. Margules & I. Musto), pp. 349–374. Proc. Workshop held at Adelaide Univ. 31 Aug. to 31 Sept. 1983, CSIRO (Aust.) Division of Water and Land Resources, Canberra, Australian Capital Territory.

Gleason, H.A. (1939) The individualistic concept of the plant association. *The American Midland Naturalist*, **21**, 92–110.

Gotelli, N.J. & Ellison, A.M. (2013) *A Primer of Ecological Statistics*, 2nd edn. Sinauer, Sunderland, Massachusetts.

Gregoire, T.G. & Valentine, H.T. (2007) *Sampling Strategies for Natural Resources and the Environment*. Chapman and Hall, London.

Grinnell, J. (1917) The niche-relationships of the California thrasher. *The Auk*, **34**, 427–433.

Huisman, J., Olff, H. & Fresco, L.F.M. (1993) A hierarchical set of models for species response analysis. *Journal of Vegetation Science*, **4**, 37–46.

Hurlbert, S.J. (1984) Pseudoreplication and design of ecological field experiments. *Ecological Monographs*, **54**, 187–211.

Ioannidis, J.P.A. (2005) Why most published research findings are false. *PLoS Medicine*, **2**, e124.

Kelly, M., Bennion, H., Burgess, A., Ellis, J., Juggins, S., Guthrie, R., Jamieson, J., Adriaenssens, V. & Yallop, M. (2009) Uncertainty in ecological status assessments of lakes and rivers using diatoms. *Hydrobiologica*, **633**, 5–15.

Kenkel, N.C., Juhász -Nagy, P. & Podani, J. (1989) On sampling procedures in population and community ecology. *Vegetatio*, **83**, 195–207.

Legendre, P., Troussellier, M., Jarry, V. & Fortin, M.-J. (1989) Design for simultaneous sampling of ecological variables: from concepts to numerical solutions. *Oikos*, **55**, 30–42.

Lohr, S.L. (2009) *Sampling: Design and Analysis*, 2nd edn. Brooks/Cole, Boston, Massachusetts, USA.

Mac Nally, R. (2000) Regression and model-building in conservation biology, biogeography and ecology: the distinction between – and reconciliation of – 'predictive' and 'explanatory' models. *Biodiversity and Conservation*, **9**, 655–671.

Matthews, T.J., Steinbauer, M.J., Tzirkalli, E., Triantis, K.A. & Whittaker, R.J. (2014) Thresholds and the species–area relationship: a synthetic analysis of habitat island datasets. *Journal of Biogeography*, **41**, 1018–1028.

Mazerolle, M.J. (2013) *AICcmodavg: model selection and multimodel inference based on (Q)AIC(c). R package version 1.35.* URL http://CRAN.R-project.org/package = AICcmodavg [accessed at 01 April 2014]

Myers, R.H., Khuri, A.I. & Carter, W.H. (1989) Response surface methodology: 1966–1988. *Technometrics*, **31**, 137–157.

Nychka, D., Furrer, R. & Sain, S. (2014) *fields: tools for spatial data. R package version 7.1.* URL: http://CRAN.R-project.org/package = fields [accessed 11 August 2014]

Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science*, **349**, aac4716.

Palmer, M.W. & White, P.S. (1994) On the existence of communities. *Journal of Vegetation Science*, **5**, 279–282.

Prinz, F., Schlange, T. & Arrowsmith, K. (2011) Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, **10**, 712.

Quinn, G.P. & Keough, M.J. (2002) *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, Cambridge.

R Development Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. URL http://www.R-project.org [accessed 10 June 2013]

Ramenskij, L.G. (1918) *Zur Methodik der Quantitativen Vegetationsaufnahme. Trudy Sowestsch.* Geobot.-lugowjedow. Moskau, Russia.

Richardson, A.D., Aubinet, M., Barr, A.G., Hollinger, D.Y., Ibrom, A., Lasslop, G. & Reichstein, M. (2012) Uncertainty quantification. *Eddy Covariance A: Practical Guide to Measurement and Data Analysis* (eds M. Aubinet, T. Vesala & D. Papale), pp. 173–209. Springer, Dordrecht, Netherlands.

Sanders, N.J. & Rahbek, C. (2012) The patterns and causes of elevational diversity gradients. *Ecography*, **35**, 1–3.

Scheiner, S.M. & Gurevitch, J. (2001) *Design and Analysis of Ecological Experiments*, 2nd edn. Oxford University Press, Oxford.

Shanks, D.R., Newell, B.R., Lee, E.H., Balakrishnan, D., Ekelund, L., Cenac, Z., Kawadia, F. & Moore, C. (2013) Priming intelligent behavior: an elusive phenomenon. *PLoS ONE*, **8**, e56515.

Simmons, J.P., Nelson, L.D. & Simonsohn, U. (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, **22**, 1359.

Southwood, T.R.E. & Henderson, P.A. (2000) *Ecological Methods*, 3rd edn. Blackwell Science, Oxford.

Stein, A. & Ettema, C. (2003) An overview of spatial sampling procedures and experimental design of spatial studies for ecosystem comparison. *Agriculture, Ecosystems and Environment*, **94**, 31–47.

Steinbauer, M.J., Dolos, K., Reinking, B. & Beierkuhnlein, C. (2012) Current measures for distance decay in similarity of species composition are influenced by study extent and grain size. *Global Ecology and Biogeography*, **21**, 1203–1212.

Taylor, J.R. (1991) *An Introduction to Error Analysis*. University Science, Sausalito, California, USA.

The Economist (2013) *Unreliable research: trouble at the lab. The Economist*, October 19th 2013. URL http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble.

Thompson, S.K. & Seber, G.A.F. (1996) *Adaptive Sampling*. Wiley, New York City, New York, USA.

Tuszynski, J. (2012) *caTools: tools: moving window statistics, GIF, Base64, ROC AUC, etc. R package version 1.14*. URL http://CRAN.R-project.org/package=caTools. [accessed 01 April 2014]

Urbanek, S. (2011) *multicore: parallel processing of R code on machines with multiple cores or CPUs. R package version 0.1-7*. URL http://CRAN.R-project.org/package=multicore [accessed 01 April 2014]

Venables, W.N. & Ripley, B.D. (2002) *Modern Applied Statistics with S*, 4th edn. Springer, New York City, New York, USA.

Whittaker, R.H. (1967) Gradient analysis of vegetation. *Biological Reviews*, **42**, 207–264.

Whittaker, R.H. (1972) Evolution and measurement of species diversity. *Taxon*, **21**, 213–251.

Received 10 August 2015; accepted 5 October 2015
Handling Editor: Matthew Schofield

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. Prediction success depicted as the precision of prediction (POP) for all five tested response patterns (A: no response, B: linear, C: exponential decay, D: unimodal centered, E: unimodal non-centered) and the average (F).

Appendix S2. Prediction success depicted as the fraction of correctly detected response types (FCR, A,C and E) as well as the precision of prediction (POP, B, D and F) for different levels of random variation of the response variable based on an average of the tested response patterns.

Appendix S3. Prediction success depicted as the fraction of correctly detected response types (FCR) for three different levels of random variation (SD = 2%, 10% and 20% of total variation) and four different levels of systematic error (SE = 0%, 5%, 10% and 20% of total variation) based on an average of the tested response patterns (without no-response).

Appendix S4. Probability of erroneous predictions depicted as the fraction of erroneous detected response types (FER) for three different levels of random variation (SD = 2%,10% and 20% of total variation) and four different levels of systematic error (SE = 0%, 5%, 10% and 20% of total variation) based on the no response pattern.

Data S1. Simulation scripts for R.