Review & synthesis

Ecography 40: 913–929, 2017
doi: 10.1111/ecog.02881

Subject Editor and Editor-in-Chief: Miguel Araújo. Accepted 7 December 2016

Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure

David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig and Carsten F. Dormann

D. R. Roberts (http://orcid.org//0000-0002-3437-2422)(droxroberts@gmail.com), S. Ciuti, S. Hauenstein, F. Hartig and C. F. Dormann, Dept of Biometry and Environmental System Analysis, Univ. of Freiburg, Freiburg, Germany. FH also at: Theoretical Ecology, Univ. of Regensburg, Regensburg, Germany. – V. Bahn, Dept of Biological Sciences, Wright State Univ., Dayton, OH, USA. – M. S. Boyce, Dept of Biological Sciences, Univ. of Alberta, Edmonton, AB, Canada. – J. Elith, G. Guillera-Arroita, J. J. Lahoz-Monfort and B. A. Wintle, School of BioSciences, Univ. of Melbourne, Melbourne, Australia. – B. Schröder, Inst. of Geoecology, Div. of Landscape Ecology and Environmental Systems Analysis, Technische Univ. Braunschweig, Braunschweig, Germany, and: Berlin-Brandenburg Inst. of Advanced Biodiversity Research (BBIB), Berlin, Germany. – W. Thuiller, Lab. d'Écologie Alpine, Univ. Joseph Fourier, Grenoble, France. – D. I. Warton, School of Mathematics and Statistics and Evolution and Ecology Research Centre, The Univ. of New South Wales, Sydney, Australia.

Ecological data often show temporal, spatial, hierarchical (random effects), or phylogenetic structure. Modern statistical approaches are increasingly accounting for such dependencies. However, when performing cross-validation, these structures are regularly ignored, resulting in serious underestimation of predictive error. One cause for the poor performance of uncorrected (random) cross-validation, noted often by modellers, are dependence structures in the data that persist as dependence structures in model residuals, violating the assumption of independence. Even more concerning, because often overlooked, is that structured data also provides ample opportunity for overfitting with non-causal predictors. This problem can persist even if remedies such as autoregressive models, generalized least squares, or mixed models are used. Block cross-validation, where data are split strategically rather than randomly, can address these issues. However, the blocking strategy must be carefully considered. Blocking in space, time, random effects or phylogenetic distance, while accounting for dependencies in the data, may also unwittingly induce extrapolations by restricting the ranges or combinations of predictor variables available for model training, thus overestimating interpolation errors. On the other hand, deliberate blocking in predictor space may also improve error estimates when extrapolation is the modelling goal. Here, we review the ecological literature on non-random and blocked cross-validation approaches. We also provide a series of simulations and case studies, in which we show that, for all instances tested, block cross-validation is nearly universally more appropriate than random cross-validation if the goal is predicting to new data or predictor space, or for selecting causal predictors. We recommend that block cross-validation be used wherever dependence structures exist in a dataset, even if no correlation structure is visible in the fitted model residuals, or if the fitted models account for such correlations.

The problem of structured data

Ecological data often show internal dependence structures: the tendency for values of nearby observations to be more similar than distant observations is widespread, if not pervasive. It can be found within every spatial scale from microhabitats to continents (spatial structure; Legendre 1993, Koenig 1999, Dormann et al. 2007), or within sequentially timed observations (temporal structure), such as in animal telemetry data (Rooney et al. 1998, Otis and White 1999) or population size estimates (Lundberg et al. 2000, Bjørnstad and Grenfell 2001). In behavioural ecology, individuals may form groups (herds, flocks, schools, packs) with synchronised activity or movement (hierarchical structure; Wu and David 2002, Sumpter 2006). In multi-species analyses or in analyses of genetic populations, evolutionary relatedness may also lead to dependence between species (phylogenetic structure) or populations of species (genetic structure) with more recent divergence will tend to be more similar than those which diverged longer ago (Harvey and Pagel 1991).

While such underlying structures in the data are not fundamentally problematic for statistical analyses, they tend to create two undesirable outcomes. First, model error, as well as neglected processes and variables connected to these structures, often leads to dependence structures in the model residuals, which violates the critical assumption of independence present in many models and methods (Legendre and Fortin 1989, Miller et al. 2007). Second, because predictor variables are often correlated with underlying dependence structures (e.g. climate with space), models may use predictors to overfit the residual dependence structure and thereby remove it, partially or completely.

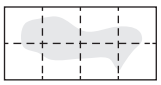
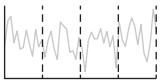
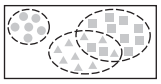
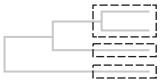
Dependence structure	Parametric solution	Blocking	Blocking illustration
Spatial	Spatial models (e.g.CAR, INLA, GWR)	Spatial	
Temporal	Time-series models (e.g.ARIMA)	Temporal	
Grouping	Mixed effect models (e.g. GLMM)	Group	
Hierarchical / Phylogenetic	Phylogenetic models (e.g. PGLS)	Hierarchical	

Figure 1. Examples of dependence structures, parametric solutions to parameter estimation, and the associated blocking approaches for cross-validation to increase reliability of prediction error estimates.

The standard statistical answer to this problem is the use of appropriate parametric models that include the respective dependence structure (Table 1), such as spatial or temporal autoregressive models, mixed models, or phylogenetic least squared regressions. In principle, these models solve the problem of independence and should allow the use of standard parametric methods for evaluating model fit and model selection (Dormann et al. 2007, Miller et al. 2007). In practice, however, specification errors as well as the problem of structural overfitting, as described above, can lead to a poor performance of these parametric model evaluations. Moreover, many popular machine-learning methods such as random forest or neural networks do not allow accounting for such dependence structures. For all these reasons, it is crucial that we have robust nonparametric methods for validation, selection, and assessment of predictive accuracy of models when used on ecological data with internal dependence structures.

Ideally, model validation, selection, and predictive errors should be calculated using independent data (Araújo et al. 2005). For example, validation may be undertaken with data from different geographic regions or spatially distinct subsets of the region, different time periods, such as historic species records from the recent past or from fossil records. Most commonly, however, either no such independent data exist or they do not meet assumptions of independence (Araújo et al. 2005). Further, changes in biological relationships, community structures, or evolutionary changes may affect species responses in different regions or time periods (Fielding and Bell 1997, Maguire et al. 2015). Because of these difficulties, predictive error on new data is commonly approximated by cross-validation, in which data are (repeatedly) split into two subsets, one used for model training and the other for model testing (see Supplementary material Appendix 1 Table A1.1 for an overview of specific approaches and Table A2 for compiled references). This principle of data splitting is central to many of today's statistical algorithms and workflows, in particular for all predictive modelling frameworks in ecology (Hastie et al. 2009). The central assumption here is that training and evaluation data are independent. If not, error estimates will be too optimistic, and model selection will favour too complex models.

Early in their development, statistical models were typically assessed on their fit to the data alone (euphemistically referred to as 'resubstitution'), representing an extreme case of non-independence of the hold-out. Of course, any such dependence of the validation with the training data will favour overfitted models (Larimore and Mehra 1985, Hawkins 2004), resulting in artificially small error estimates and thus overly optimistic estimates of model performance (Mosteller and Tukey 1977, Olden et al. 2002, Arlot and Celisse 2010). A similar situation occurs when there are dependence structures in the data. When data held-out for validation are drawn from nearby in the dependence structure (e.g. close in space or time, from the same herd, etc.) the independence of evaluation data can be compromised (Dormann et al. 2007, Kuhn 2007, Hastie et al. 2009, Telford and Birks 2009, Bahn and McGill 2013), again producing overly optimistic estimates of prediction error (Mosteller and Tukey 1977, Picard and Cook 1984), and potentially leading to erroneous scientific conclusions (Kuhn 2007, Hastie et al. 2009, Telford and Birks 2009).

In other words, non-independence of hold-out data from the training data erroneously makes models appear more reliable than they are, enticing us to have more faith in their predictions than is actually warranted. Comparative studies of model validations for ecological applications have consistently demonstrated this e.g. (Olden et al. 2002, Reineking and Schröder 2003, Araújo et al. 2005, Veloz 2009, Lieske and Bender 2011, Roberts and Hamann 2012a, Wenger and Olden 2012, Bahn and McGill 2013). Problematically, modellers often partook (and frequently still do) in what Stone (1974, p. 111) labels "controlled division" of data, wherein "the cautious statistician … sets aside a *randomly* [our emphasis] selected part of his sample without looking at it and then plays without inhibition with what's left, confident in the knowledge that the set-aside data

Table 1. Guidelines for achieving reliable error estimates in consideration of modelling objectives (extrapolation vs. interpolation) and cross-validation approaches that may block in predictor space, structure, both predictor space and structure, or neither.

		Cross-validation structure	
		Random	Blocked
Cross-validation predictor space	Random	Correct interpolation error without random structure	Correct interpolation error with random structure
	Blocked	Correct extrapolation error without random structure	Correct extrapolation error with random structure

will deliver an unbiased judgment on the efficacy of his analysis." Of course, such random data splitting does not provide independent validation when a dependence structure is present and, thus, "unbiased judgment" is compromised.

In response, statisticians have introduced a smorgasbord of cross-validation approaches in an effort to achieve unbiased error and parameter estimates (Stone 1974, Picard and Cook 1984, Shao 1993, Kohavi 1995), many of which have been incorporated into ecological studies (Mankin et al. 1977, Verbyla and Litvaitis 1989, Power 1993, Rykiel 1996). Early solutions were leave-*n*-out cross-validation approaches (Stone 1974, Picard and Cook 1984) that run iteratively, each time withholding a small randomly selected subset of the data for testing. Because these approaches have also been shown to produce biased error estimates (Shao 1993, Kohavi 1995, Telford et al. 2004, Amesbury et al. 2013), further corrections have been proposed, for example by incorporating distance-based buffers around hold-out points (Bahn 2009, Telford and Birks 2009, Le Rest et al. 2014).

A general strategy to increase independence in cross-validation is to split data into 'blocks' at some central point(s) of the dependence structure, such as in time or space. There are some examples of block cross-validations in the ecological literature, implemented with a wide variety of stated objectives: most often for identifying non-transferability or the general inability to extrapolate, but also for increasing independence, avoiding overfitting, providing more reliable error estimates, or selecting better predictive models (Supplementary material Appendix 1 Table A1.2–A1.3). When systematically compared with random data splits, they consistently demonstrate larger errors in predictions (Burman et al. 1994, Arlot and Celisse 2010, Lieske and Bender 2011, Roberts and Hamann 2012a, Wenger and Olden 2012, Bahn and McGill 2013, Radosavljevic and Anderson 2014). It should be noted, however, that few studies have explicitly demonstrated that the estimates resulting from blocked cross-validations are indeed closer to the 'true' error that would be expected for a truly independent dataset (but see Trachsel and Telford 2016).

However, there is also reason to be cautious about reported block cross-validation errors. While block cross-validation addresses correlations, it can create a new validation problem: if blocking structures follow environmental gradients, blocking may hold out entire portions of the predictor space (i.e. ranges and/or combinations of predictor variables), introducing extrapolation between cross-validation folds (Kennard and Stone 1969, Snee 1977). Consequently, when predicting to the hold-out data, the model has to predict outside the ranges or into new combinations of predictor values of those included in the training folds. This could occur, for example, with spatial data splits, as climatic environments tend to be geographically structured (e.g. latitudinal gradients of temperature), or in temporal splits, as some periods will not have experienced certain predictor conditions (Zurell et al. 2012). In some cases, one may make a virtue of necessity, using this to test a model's extrapolation error. In general, however, the concern remains that unwanted blocking of the environmental space could lead to an overestimation of interpolation errors.

Our objective for this article is to examine the utility and application of block cross-validation. We review existing approaches, clarify the reasons for their use and their potential implementations, and discuss their shortcomings and challenges. We believe a better understanding of blocking and its relevance is highly important. Currently, blocking is not widely used in biogeographical studies and, when it is, the motivations for doing so are often not clear. This is a concern when so many studies now involve prediction to new times and/or places. However, we also demonstrate compelling reasons for using block cross-validation even for model predictions to the same time and same region. Moreover, the majority of applications in our review come from the species distribution modelling literature, block cross-validation has broad applicability to virtually any ecological analysis performed on structured data. We illustrate this point through simulations and case studies across a range of ecological questions. Specifically, we look at four block cross-validation scenarios: spatial blocking, blocking by hierarchical groups, phylogenetic blocking, and blocking in predictor space. Via these analyses, we demonstrate that:

1) random cross-validations, even with models that should correct for dependence structures, yield error estimates that are too low;
2) block cross-validation does not only increase error estimates but actually provides estimates that are closer to true values;
3) blocking in structured data often restricts the predictor space and controlling this tendency may be necessary, depending on whether inter or extrapolation in predictor space is the goal.

Cross-validation with structured data

Ecological variables (observations of biota) commonly contain four types of internal structure: autocorrelation in time, autocorrelation in space, group dependence structures, and phylogenetic structure (i.e. relatedness). These can lead to two issues in statistical models: 1) non-independence of residuals, and 2) overfitting to the dependence structure of the data. The first issue arises, for example, when a model misses a structured variable, or if it does not describe its effect on the response perfectly. Non-independence of residuals violates a central assumption of regression models and other statistical methods, typically leading to over-optimistic confidence intervals and incorrect p-values (Ives and Zhu 2006). The second issue, overfitting to the dependence structure of the data, describes the phenomenon that the model may absorb structured residual variation with another predictor (e.g. time or space themselves or another covariate that correlates with them). This can mask both the first problem and the underlying model misspecification, creating an overfitted model that does not predict well to new data.

To clarify these two issues, we provide four ecological examples where both issues could emerge:

1) *Temporal structure* - Imagine that we have annual time series data of antelope population size, which fluctuates over time but always tends to be similar to a previous year's size. Also imagine that we have annual rainfall in each relevant year as a covariate. Population size may be partly driven by rainfall but is also affected by other covariates which we have not measured but that may also be structured in time, leading to model residuals that are temporally autocorrelated (non-independence of residuals). However, because the missing covariates are also likely to correlate with rainfall, and that they all follow similar temporal structure, the model may attribute part of the effect of these other covariates to rainfall itself, which would result in biased parameter estimates and reduced autocorrelation of residuals (overfitting). Rainfall and demographically induced temporal autocorrelation are confounded.

2) *Spatial structure* - Imagine the distribution of an anolis lizard across an island archipelago, which likely dispersed gradually throughout the individual islands from a single source, so their populations are spatially structured (i.e. data from nearby islands are likely to be more similar). If we model lizard distribution with climate, we will certainly end up with spatial autocorrelation in model errors due to the historic dispersal pattern of populations. However, these residuals will be reduced because we also certainly (and unintentionally) alias part of the geographic space via spatially structured environments. Thus, even if climate was immaterial to the species, it would be used by the model as a trend-surface-regression to reduce residual spatial autocorrelation (overfitting). Geographic space and climate space are confounded.

3) *Hierarchical or group structure* - Imagine we have recorded observations of hyena movements paired with tree cover classes. While hyena movement may be partly driven by tree cover, they are also driven by movements of other hyena individuals in the same cackle. A habitat selection model based on tree cover will then result in residuals autocorrelated by individual animals or even by groups themselves (non-independence of residuals). Further, each cackle likely moves within different tree covers, particularly if the cackles tend to avoid one-another (i.e. tree cover correlates with individuals or groups). Therefore, other variables that correlate with individuals or groups, and thus also with tree cover, may be partly accounted for in the model by tree cover itself (overfitting), further reducing model residuals. We are unable to separate the contribution to the model of tree cover and the underlying random structure that tree cover represents, thus confounding individual or grouping structures.

4) *Phylogenetic or genetic structure* - Imagine that we have drought tolerance data for a tree species, which tends to vary across several genotypes. Also imagine that we use distance from the coastline as a covariate, knowing that interior populations can survive drier conditions. Drought tolerance may be partly driven by the distance from the coast, but is also affected by other, unmeasured covariates, leading to model residuals that are autocorrelated by phylogenetic relatedness (non-independence of residuals). It is possible that these missing covariates correlate both with coastal distance (say, if they are environmentally driven) and with genetic relatedness (especially if the species migrated post-glacially from a single ice-age refugium, structuring genetic relatedness in space). In this case, the model may attribute part of the effect of the unmeasured covariates to coastal distance, which would result in biased parameter estimates and reduced autocorrelation of residuals (overfitting). Coastal distance is confounded with phylogenetic dissimilarity. The same type of situation could apply when considering multiple species that are phylogenetically related.

Non-independence of residuals may be addressed by explicitly modelling correlation structures, such as with autoregressive models (for space and/or time), hierarchical models (for describing nested structures), or phylogenetic contrasts (Ives and Zhu 2006; Table 1). For example, in parametric models, such problems are addressed by moving from simple regression models with independent random error assumptions to more complex model structures such as conditional spatial autoregressive models (CAR; Cressie 1993), integrated nested Laplace approximations (INLA; Rue et al. 2009), or geographically weighted regressions (GWR; Fotheringham et al. 2002), time series methods such as autoregressive integrated moving averages (ARIMA; Brockwell and Davis 1996), methods that include random effects such as generalised linear mixed models (GLMM; Breslow and Clayton 1993), or phylogenetic comparative methods such as phylogenetic generalised least squares (PGLS; Grafen 1989). These model structures can account for correlations among data points, yielding unbiased estimates, at least in theory.

Overfitting is a more insidious problem because it can easily escape detection unless cross-validations are carefully implemented. Here, structure in observations (e.g. space, time, groups) is being explained by the model through some other non-causal covariate. This is particularly common in ecology because covariates themselves can be structured in the same way as the residuals (i.e. in space, time, phylogeny, etc.). Thus, covariates need not be orthogonal to model structure, as assumed implicitly by methods in the previous paragraph. Resulting model predictions may perform fine in a situation where the correlation structure between non-causal and the "true" predictors (i.e. underlying structures) remains unchanged (Bahn and McGill 2007), but they could completely fail when predicting to novel situations. Methods that directly target residuals (e.g. spatial variograms, or regression models with structured errors; Table 1) may fail to detect this problem because overfitting may hide the structure of the residuals. This can occur even when using models that account for dependence structures, such as those discussed above (and listed in Table 1).

As a consequence of both problems, cross-validation on random data splits (all of which will be largely consistent in underlying structure) as well as the various parametric modelling options (Table 1) will tend to underestimate predictive error (Araújo et al. 2005, Veloz 2009, Bahn and McGill 2013), leading to false confidence in model predictions. We show in our later examples that this problem persists, although to a lower extent, even if appropriate parametric models (e.g. spatial models, mixed models, PGLS) are used. To address these issues associated with dependence structures, whether or not we know they are

present, we can introduce blocking across the given correlation structure into our cross-validations (Table 1). Different modelling objectives and different underlying data structures will necessitate different blocking approaches. When models are intended only to predict within the same spatial and temporal ranges or on the same individuals or groups by which they were trained, random cross-validation may yield fair error estimates because the model's conditionals do not change. When models are intended to only to predict on the same data structure, without the desire to make causal inferences, random cross-validation may yield fair error estimates. However, if the goal is to infer causal predictors, or predict into new dependence structures (i.e. new locations, new time periods, new individuals or groups, or for new species within a phylogeny), blocking is required. Moreover, blocking can also be used to estimate errors under extrapolation in predictor space, which will be discussed in the next section.

Blocking to account for spatial and temporal autocorrelation

When validation data are randomly selected for cross-validation from the entire spatial domain, training and validation data from nearby locations will be dependent (spatial autocorrelation). Consequently, if the objective is to project outside the spatial structure of the training data, error estimates from random cross-validations will be overly optimistic. To address this, blocks can be designed across the spatial structure itself (i.e. in contiguous geographic space). This effectively forces testing on more spatially distant records, thus decreasing spatial dependence and reducing optimism in error estimates (Trachsel and Telford 2016). We demonstrate this via a simulation in Box 1. Temporal autocorrelation, which is functionally similar to spatial autocorrelation in a single dimension, presents the same dependence challenges. When models are intended to predict in time, blocks may be drawn in the same manner (i.e. blocks of contiguous time) to better ensure independence between cross-validation folds (Burman et al. 1994, Racine 2000, Bergmeir and Benitez 2012).

Blocking to account for random effect structures

A somewhat different structure is presented by hierarchical data, such as blocked or nested experimental designs, data replicated by individuals in groups, or repeated measurements such as animal telemetry data. In these cases, data are structured by units such that observations within the same block, individual, or group will tend be more similar in quality and more dependent in response. Just as with spatial or temporal autocorrelation, models parameterised on such data may be fitted to the grouping structure itself via covariate predictors and consequent optimism in error estimates would be expected from random cross-validations (i.e. that predict only on the same grouping structure) relative to the error expected when predicting to new individuals or groups.

We present an example in Box 2, in which we estimate resource selection functions (RSFs; Manly et al. 2002) with repeated movement observations of individual ungulates. In this case study, blocking for cross-validation by individual animals circumvents the problem of the underlying random structure and delivers a more realistic error estimate for predictions onto new individuals. This case study also illustrates that, while a including random effects in a regression approach (i.e. a mixed model) might yield unbiased model parameter estimates, it cannot offer a reliable uncertainty for those estimates and, further, does not address the problem of underestimation of predictive error from cross-validations with random data splits.

Blocking to account for phylogenetic correlations

Species properties are often phylogenetically conserved, meaning that closely related species tend to be more similar to each other than distant relatives. Consequently, analyzing data across species can lead to phylogenetically correlated residuals, resulting in individual observations that are not independent (Felsenstein 1985). Just as in time, space, or individuals and groups, phylogenetic structure can be overfit by the model when covariates correlate with phylogenetic structure. It has therefore become common in ecological analysis to fit regression models that include phylogenetic structure in their residuals (PGLS; Table 1; Revell 2010). To ensure independence in cross-validation, it may also be necessary to block observations by phylogenetic distance. To our knowledge, such an approach to cross-validation has not been undertaken in the phylogenetic literature. We demonstrate in Box 3 that this greatly improves inference for phylogenetically structured data.

Disentangling blocking and extrapolation

So far, we have discussed block cross-validation as a means for model selection (Box 3) and calculating a corrected interpolation error in the presence of correlation structures within data or residuals (Box 1, 2). Blocking, however, can also be used to estimate extrapolation error. Extrapolations into new predictor space are different from changes in underlying structure of the data: the latter only changes correlations between predictors, while the former requires the model to predict a response in an area of predictor space about which they are uninformed. Models typically show larger error when extrapolating into these no-analogue conditions (Pearson 2006, Elith and Graham 2009). Consequently, if our modelling goal is extrapolation, we are likely to underestimate prediction errors with standard cross-validation approaches (Heikkinen et al. 2012). On the other hand, blocking may inadvertently restrict the predictor space for model training, especially as data structures are often collinear with clines in predictor variables (e.g. spatial temperature clines), creating overly pessimistic error estimates when model extrapolation is of no interest. Thus, when making decisions about cross-validation approaches, model objectives must be carefully considered (Table 1).

Box 1. Spatial blocking

Spatial structure in data can lead to the underestimation of model prediction error when covariate predictors allow models to fit these structural patterns. In this simulation we investigate the ability of spatial blocking strategies to minimize this problem. We simulated data of species abundances on a 50 × 50 grid that depended in complex ways (interactions, non-linear combinations, limiting effects, and exclusion by disease) on four spatially autocorrelated 'environmental' variables. We modelled the data using Random Forest (Breiman 2001) and compared the root mean squared errors (RMSE) among evaluation strategies. The results are based on 100 replicated landscapes (Supplementary material Appendix 2 for details).

Evaluation strategies tested included 1) using the same data for training and evaluation (resubstitution), 2) randomly splitting data into training and test data (random), 3) splitting the data into training and test data blocked in space with block sizes 10 × 10, 20 × 20 cells and half of the grid (25 × 50 cells), and 4) a leave-one-out cross validation (LOO) with spatial buffering , in which the cell held-out for evaluation is buffered by a circle of cells (radius 5, 8 or 10), which are also omitted from the training set. We either used all test sites resulting from the evaluation strategies, even if they were environmentally non-analogous to training data (no-analogues included), or restricted testing sites to analogue ones (minimal environmental extrapolation). Cross-validations were compared to an 'ideal' RMSE, which was estimated by producing a model for each of the 100 landscapes and predicting to the remaining 99 then averaging the errors to achieve a single RMSE per landscape.

Our results show that ignoring dependence between training and test sites (resubstitution and randomly drawn folds) lead to artificially low error estimates, while block cross-validation and the buffered LOO produce error estimates much closer to the true error as determined by predicting on new, independent data, particularly when test sites are forced to be environmentally analogous to training sites. We also find that the size of the blocks needs to be substantially larger than the range of the spatial autocorrelation in the model residuals (~10 units) to provide a good error estimate, while a buffer size equivalent to distances at which residual autocorrelation is reduced to zero suffices for the buffered LOO (Supplementary material Appendix 2 Fig. A1.1).

Complete R scripts for this simulation are provided in Supplementary material Appendix 6.

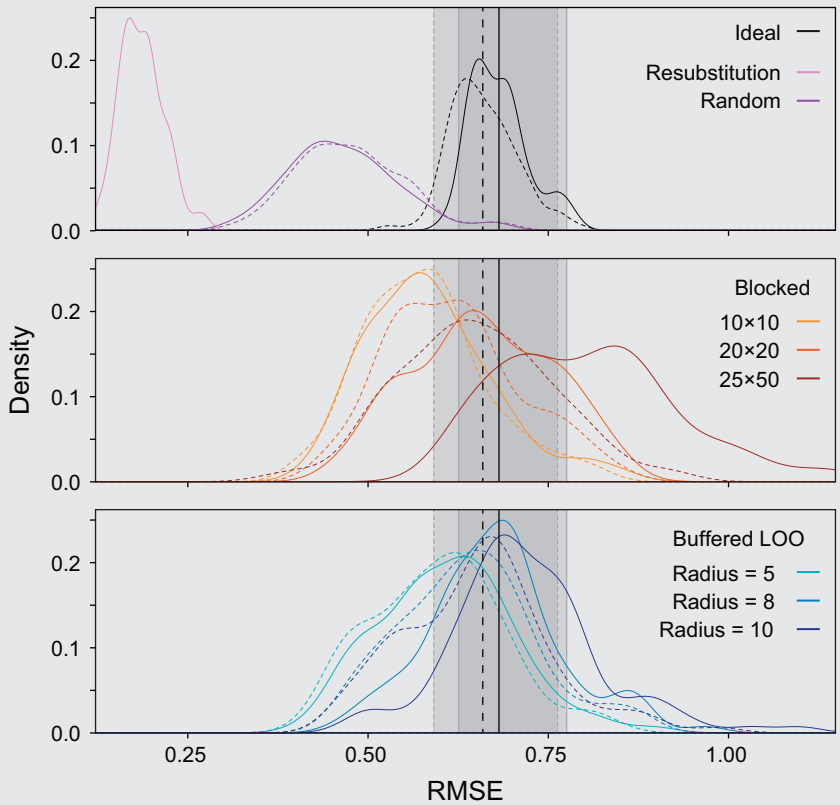


Figure B1.1. Root mean squared error (RMSE, n = 100 simulations) for the various spatially blocked cross-validation approaches. Semi-transparent shading and black vertical lines respectively show the 95% range and mean values of true RMSE (Ideal), which is determined by predicting onto independent realizations of the simulation. Solid lines show RMSE distributions with all test locations included while dashed lines show RMSE distributions with test locations non-analogous to training locations removed.

Box 2. Blocking by individual or group

We estimated resource selection functions (RSFs) for 43 female elk *Cervus elaphus* monitored using satellite telemetry in Alberta, Canada. We fitted generalized linear mixed models (GLMM) with a Bernoulli response (1 = use by elk; 0 = available, i.e. random points drawn from elk home ranges), environmental covariates as predictors, and elk individual as a random intercept. Exponentiated non-intercept parameters estimates of the GLMMs were interpreted as relative selection strength in favour of any given predictor (Lele et al. 2013).

RSFs were evaluated using five-fold cross-validation as proposed by Boyce et al. (2002), based on nonparametric correlation between RSF bins and area-adjusted frequencies for each withheld sub-sample of the data in turn. A model with good predictive performance has a strong positive correlation (Boyce et al. 2002). We evaluated the performance of a full resubstitution model (train data = test data) and three five-fold cross-validations, each with a different blocking design: 1) random, in which each elk contributed to each fold with 20% of its position fixes (no blocking); 2) randomly selected individuals, in which each elk contributed to one fold with 100% of its fixes and home ranges of elk belonging to different folds may overlap (blocked by individual); and 3) spatially blocked individuals, in which each elk contributed to one fold with 100% of its fixes and selected in such a way that home ranges of elk belonging to different folds do not overlap (blocked by individuals that behave independently). Extended methods and complete results are provided in Supplementary material Appendix 3.

Evaluation by both resubstitution and random cross-validation erroneously suggests outstanding model performance (Fig. B2.1a–b). In contrast, both blocked cross-validations (by individual and by spatially blocked individual) showed notably lower performances on average and much higher variability in Spearman-rank correlations across folds (Fig. B2.1ab). Cross-validation blocked by random individuals resulted in a notable decrease in model evaluation performance relative to random splits across all position fixes. Blocking by spatially independent individuals resulted in no further decrease in model performance, suggesting that independence between folds was achieved at the level of individual animals (or, ecologically speaking: individuals with overlapping home ranges did not behave more similarly than any two random individuals). Parameter estimates, while consistent on average across methods, covered a wider breadth of values in the blocked cross-validations, providing a measure of uncertainty for their true values (Supplementary material Appendix 3).

Both the acceptance of precision in parameter estimates as well as optimism in model validations due to non-independent folds are prevalent in RSFs studies (Supplementary material Appendix 1 Table A1.2–A1.3; but see Wiens et al. 2008, Koper and Manseau 2009, Coe et al. 2011). Block cross-validation can help avoid such overconfidence in model performance and foster greater care in the search for sound model structures.

Complete R scripts and data for this case study are provided in Supplementary material Appendix 6.

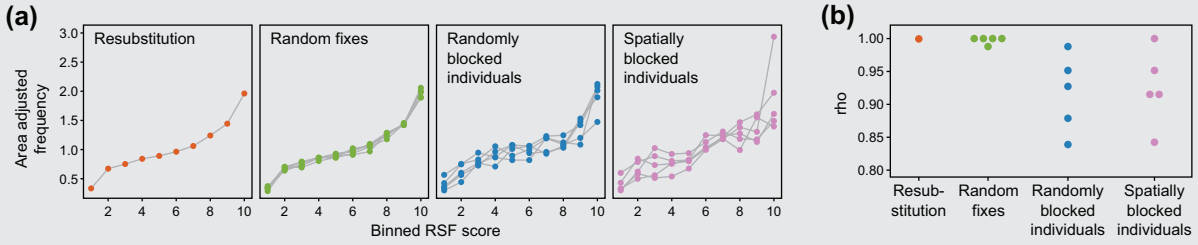


Figure B2.1. Summaries of resource selection function (RSF) implementations incorporating several validation approaches including resubstitution, cross-validation with random fixes, cross-validation with randomly blocked individuals, and cross-validation with spatially blocked individuals, showing (a) the area-adjusted frequency of RSF score bins and (b) Spearman's rank correlations (rho) between RSF bin ranks and area-adjusted frequencies for each cross-validation fold.

Avoiding extrapolation

Environments tend to be structured in space and time: climates tend to be similar in nearby locations just as they tend to be similar in consecutive time periods. Therefore, because blocking to achieve structural independence in cross-validation requires the grouping of similar structural groups (e.g. contiguous space or time), such blocking also might group similar predictor space together, potentially removing all such like space from the remaining data. This effect is likely to be more pronounced when simple sets of predictor variables with global effects such as climate are used, in contrast to variables that explain distributions at finer grains and more local extents (Mackey and Lindenmayer 2001). When models are intended to interpolate (i.e. predict only to similar predictor space), blocking may induce extrapolation in cross-validation when unnecessary. While it's unlikely that this can be avoided entirely, these effects can be minimised by 1) using blocks no larger than necessary considering the grain and extent of analysis and the spatial scale of patterning of environment, 2) using as much data as possible for model training, and 3) representing predictors equally across blocks or folds.

Extrapolation will generally increase as more data are withheld for testing (Box 4). Consequently, predictor coverage in training data can be maximised by making blocks

(or leave-one-out buffers) only as large as required. The minimum blocking distance should be the extent of autocorrelation in model residuals ('autocorrelation requirements' in Fig. 2a). If correlation structures are multidimensional, anisotropic analyses of residual autocorrelation (in individual dimensions) may also allow blocks to be narrower in one direction than the other, while still achieving independence. For example, a model describing spatial data that is missing a key temperature variable may result in residual autocorrelation that extends in the north–south direction (the general direction of the temperature gradient) much farther than in the east–west direction. Dividing such data into square blocks or defining a circular buffer radius for a leave-one-out approach based on an isotropic autocorrelation distance would, in such a case, result in unnecessarily large east–west block distances, and potentially introduce unnecessary extrapolation into the cross-validation. Portrait-oriented rectangular blocks might better limit extrapolation.

We state the extents of residual autocorrelation as a blocking minimum because, as explained above, overfitting via structural covariates may reduce residual autocorrelation while not offering increased independence. Therefore, larger blocks than suggested by variograms or other measures of autocorrelation may be required to avoid optimistic error estimates, though the extent of this effect is unlikely to be known by the modeller. While the potential for introducing extrapolation is higher when blocks are made conservatively large, this can be mitigated through different approaches to block fold assignments (Fig. 2b).

While the preferred number of folds in *k*-fold approaches has been suggested to be between 5 and 10 (Kohavi 1995, Hastie et al. 2009), such recommendations are perhaps more appropriate for random splitting where an ad hoc number of folds must be chosen. To include as much data for model training as possible in each cross-validation iteration, each block should be its own fold. This, of course, maximises required iterations of the cross-validation, resulting in a potentially computationally expensive procedure, particularly when models are slow to fit or when data sets are very large ('computational limits' in Fig. 2a). However, while a

Box 3. Blocking to address phylogenetic correlation

To show the use of phylogenetic blocking, we simulate a simple trait-environment relationship (body mass versus latitude) with residual variation structured by phylogenetic distance, then cross-validate regression predictions using three approaches: a random *k*-fold, a *k*-fold blocked in phylogenetic distance, and a leave-one-out approach with buffering by phylogenetic distance. We also use these cross-validations for model selection, as well as considering model selection based on AIC of a standard regression (LM) and a geographic least squares regression (GLS).

Trait-environment data were random environmental observations (latitudes between 0 and 90) for 50 hypothetical species with a phylogeny created by a standard birth-death process. Body mass response was calculated from a quadratic (3 parameter) function, to which we added phylogenetically structured error by sampling from a multivariate normal distribution with phylogenetic distance as covariance (Fig. B3.1a–b). Semivariograms indicated that residual autocorrelation did not extend beyond ~0.5 units of phylogenetic distance.

Model selection was between eight regressions of increasing complexity (i.e. increasing polynomial order), based on minimum AIC for the LM and GLS resubstitution approaches or based on minimum root mean squared error (RMSE) for the cross-validations. Three cross-validation approaches were considered. First, we ran 5- and 10-fold cross-validations with data assigned to folds randomly. Second, we ran blocked 5- and 10-fold cross-validations with folds defined by hierarchical clustering of the phylogenetic distances (Fig. B3.1b). Last, we implemented a phylogenetically independent leave-one-out cross-validation, in which each data point was withheld in turn for model testing and the remaining data used for model training, with the exception of data within a predetermined buffer of phylogenetic distance (either 0.00, 0.25, 0.50, 0.75 or 1.00 phylogenetic distance units) around the withheld point. The entire data building and cross-validation simulation process was repeated 100 times. Extended methods and complete results are provided in Supplementary material Appendix 4.

For model selection (Fig. B3.3c), the GLS was the best model selection tool of any approach (correct structure in 60% of the simulations), while the blocked cross-validations and buffered leave-one-out approaches also performed well. The LM, the random cross-validations, and the unbuffered leave-one-out were the worst (correct structure in 12–18% of simulations), more often choosing overly complex models. For error estimates, blocked and leave-one-out cross-validations resulted in both median RMSE and ranges of RMSE better-approximating the true errors in the data generating model, while both the LM and GLS resubstitution as well as the random cross-validations and leave-one-out cross-validations with smaller buffers gave optimistic error estimates (Fig. B3.1d). Only the five-fold blocked cross-validation and the leave-one-out with the largest buffer size (1.0) resulted in RMSE values higher than those of the true model (i.e. overly-pessimistic validations).

Generally speaking, GLS reduced overfitting in model selection compared to the non-independent approaches (i.e. LM, random cross-validations, and leave-one-out cross-validations with smaller buffer sizes). However, error estimates for GLS, while an improvement on the non-independent approaches, were still optimistic. The block cross-validations and leave-one-out cross-validations with sufficiently large buffers provided the best combination of model selection and reliable error estimation.

Complete R scripts for this simulation are provided in Supplementary material Appendix 6.

(Continued)

Box 3. (Continued)

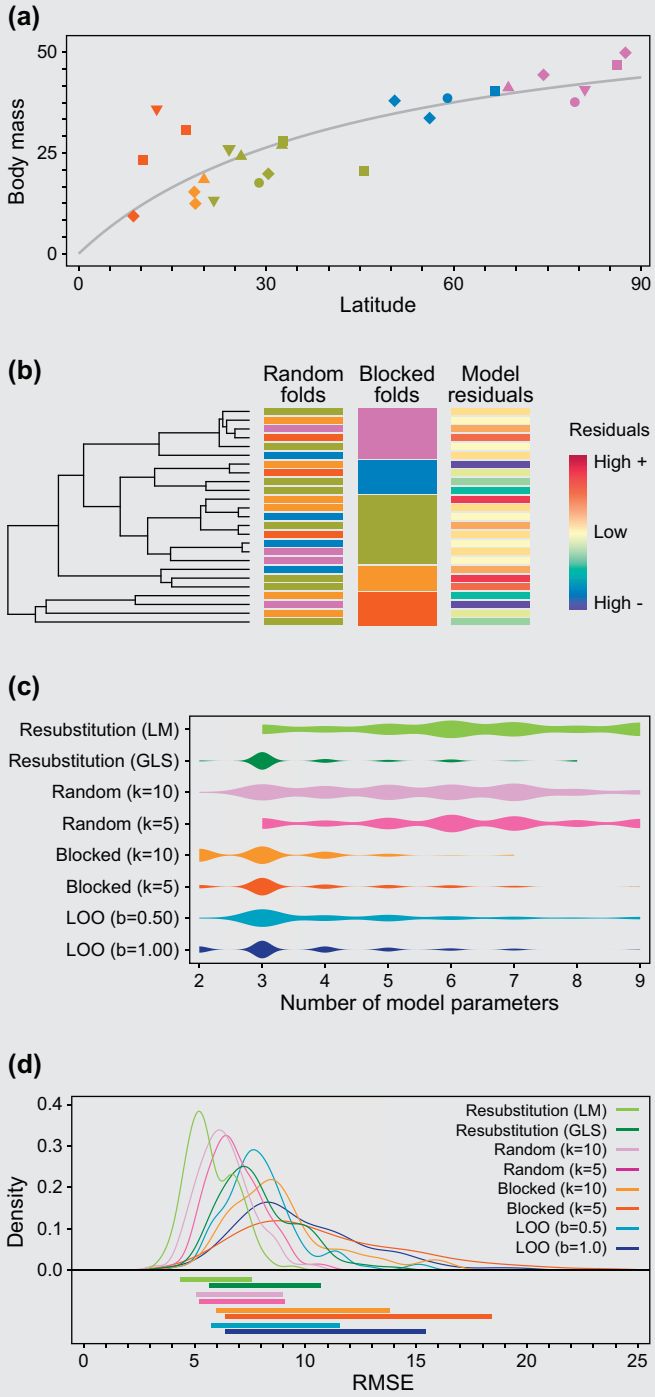


Figure B3.1. Blocking in phylogenetic space. (a) Sample regression (grey line) showing the simulated relationship between a species trait (body mass) and the environment (latitude) with phylogenetically structured correlation in the residuals, conforming to the assumptions of a PGLS model. Shapes represent random cross-validation folds and colours represent blocked cross-validation folds. (b) Sample phylogenetic tree in which 25 tips (species) are assigned to cross-validation folds randomly or assigned to folds by hypothetical clades based on phylogenetic distance. Autocorrelation in the phylogenetic structure can be visualised as simulated trait residuals structured by genetic distance in the tree. (c) Violin plot showing the distribution of model complexity (number of parameters in the selected model) for each cross-validation approach across the simulations. The number of parameters in the true data generating model (3 parameters) is shaded in grey. (d) Results of the phylogenetic blocking simulation showing distributions of RMSE in trait predictions from the resubstitution, from both the random and blocked cross-validations (with the number of folds, *k*, shown in brackets), and from the buffered leave-one-out (LOO) cross-validation (with the buffer size, *b*, shown in brackets). The shaded grey area represents the 5th to 95 percentile range of the true model RMSE. Horizontal bars below the plot show the 5th to 95th percentile range of the RMSE for each approach.

Box 4. Blocking for extrapolation

We examined the effect of blocking in environment on cross-validation, in a typical species distribution modelling approach for Douglas-fir *Pseudotsuga menziesii* habitats in western North America. Species presence–absence records were paired with climate data from the 1961–1990 period and divided into groups for *k*-fold cross-validation using several data splitting approaches: random splits, splits in geographic space, splits in predictor space, as well as data resubstitution (no splitting). Geographic splits were implemented with a two-fold checkerboard pattern across spatial grids of varying size. We also implemented a buffered leave-one-out cross-validation with buffer radii of 100, 500, 1000 and 1500 km (Box 1).

Correlograms indicated that residual autocorrelation was virtually non-existent past distances of ~220 km. Modelling was done with Random Forest (Breiman 2001) and models were evaluated on predictions to all folds using AUC. Extrapolation was quantified by 1) calculating the Euclidean distance across all principal component predictors from each point in the withheld fold back to each point in the model training data, 2) for each withheld point, calculating the first percentile distance, and 3) calculating the average of the first percentile distances from all points in the withheld data. See Supplementary material Appendix 5 for detailed methods and results.

Both spatial and environmental blocking induce extrapolation between folds (Fig. B4.1a). The largest spatial blocks (e.g. 1 × 2 blocking) and the coarsest environmental blocks (e.g. 2-group cluster analysis or splitting only in PC1) result in both the largest environmental distances and the lowest estimates of predictive accuracy (AUC), with the effect being much stronger for spatial blocks than for purely environmental blocks (Fig. B4.1a). There is a small but visible decrease in AUC for environmental blocks relative to spatial blocks at the same geographic distance, suggesting a cumulative effect on predictive accuracy when spatial and environmental extrapolation are combined (Fig. B4.1b). The buffered leave-one-out approach both minimises extrapolation and increases predictive accuracy relative to other methods at similar geographic distances.

While the effect of spatial autocorrelation may account for decreasing accuracy at distances up to ~220 km, it cannot explain the continued decrease in AUC at much larger distances. Also, with only a moderate effect on accuracy, environmental extrapolation requirements are also unlikely to be the cause of this decrease. More likely, across larger spatial blocks, the underlying spatial structure changes (e.g. competition regimes, disease presences, changes in local adaptations of genetic populations, etc.). While some of this structure may be overfit with spatially autocorrelated predictors, this overfit is likely to break down across space, thus reducing model predictive accuracy in new regions (i.e. into alternative blocks). This overfit also hides these effects from our measurements of spatial autocorrelation, as correlograms or semivariograms were built using residuals from a full data model that could be overfit to the complete spatial structure.

In summary, while purely environmental splits force extrapolation, they are unlikely to account for spatial autocorrelation or structural overfitting because blocks may be spatially intertwined. Further, smaller spatial blocks, even those that seemingly account for residual autocorrelation, may be insufficiently large to account for structural overfitting.

Complete R scripts and data for this case study are provided in Supplementary material Appendix 6.

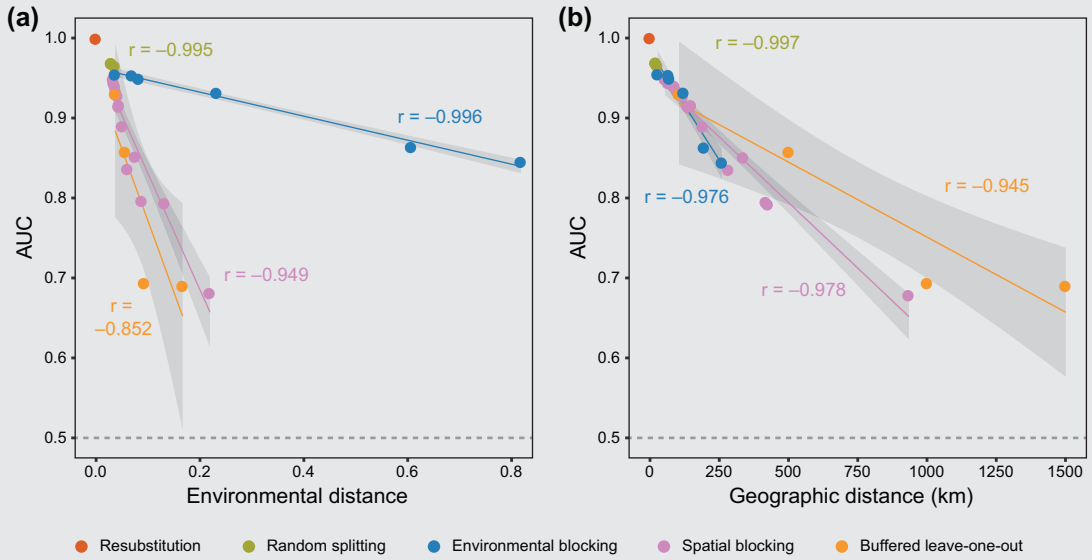


Figure B4.1. Model prediction accuracy (AUC) as function of the minimum (a) environmental distance and (b) minimum geographic distance between training and test data in various *k*-fold blocking approaches. While relationships are drawn as linear, the theoretical minimum AUC is 0.5 (for a random model).

cross-validation with a large number of folds might be computationally intensive, there is no conceptual barrier to it so long as validation data meet the assumptions of independence. That said, with small values of *k*, resulting in limited model training data ('data limits' in Fig. 2a), bias in cross-validation folds may become problematic, so the value of *k* may depend strongly on the overall data quantity (Hastie et al. 2009). The recycling of training data is, of course, maximised in leave-one-out approaches, which are also the most computationally intensive, requiring a new model to be fitted for each point in the data.

When leave-one-out approaches or *k*-fold approaches with numerous folds are not feasible or not desired, assignments of numerous blocks to fewer folds can be implemented in ways that ensure a greater variability of predictors are represented in each fold (Fig. 2b). While random assignment of blocks to folds might result in good representation of predictor space in all folds, other approaches may better ensure this. For example, blocks can be systematically assigned to folds in checkerboard or repeating patterns to distribute them evenly across the data (Fig. 2b, 'systematic'). A more directed approach could also be to divide predictors manually between folds (e.g. manually distribute blocks with similar values for key environmental values between folds). For more complex predictor space (i.e. more variables), random fold assignments can be repeated many times, measuring predictor dissimilarity between folds for each iteration, then choosing the optimal assignment resulting in lowest dissimilarity between folds (Fig. 2f, 'optimised random'). This approach, of course, only ensures that predictor space in each fold will be equally different and not necessarily different in the same way. We note that whilst this is called 'block' cross-validation, the implication is not that the divisions need to be rectangular. For instance, blocks might be based on sampling units themselves (Buston and Elith 2011), Box 2, or on distinct geographic features such as river catchments (Chee and Elith 2012) or mountain ranges (Bulluck et al. 2006).

Deliberately inducing extrapolation

In practice, modellers are often not only interested in the accuracy of their model predictions within the domain for which data exists (interpolation), but also beyond this domain (extrapolation). The need for such estimates is apparent in applications such as species habitat projections under future climate change, for which the prediction data is likely to contain no-analogue predictor space, i.e. conditions not observed within the training data (Williams and Jackson 2007). Such extrapolation requirements are relatively straightforward to identify and measure via comparisons of training and prediction data, such as by examining individual variable ranges or by measurements of multivariate distances such as the MESS maps in Maxent and related procedures (Elith et al. 2010, Zurell et al. 2012, Mesgaran et al. 2014).

The key question for model extrapolation then is not whether a model is still 'valid' when applied to new data (it almost certainly is not), but rather to what degree the violation of assumptions undermines predictive accuracy. Extrapolation errors are difficult to estimate because no data exist in the domain to which the model is predicting. In such cases, we may consider cross-validation strategies that try to simulate model extrapolation: splitting training and testing data so that the domain of predictor combinations in both sets is not overlapping. To sensibly interpret the results, we require some measure of dissimilarity in predictor space, a metric not completely straightforward to quantify. Models may include numerous predictors, some of which are autocorrelated and not all of which are equally important to every species. The simplest metrics of dissimilarity are comparisons of individual variable ranges (Capinha et al. 2012, Anderson 2013) that, while identifying extreme values in single variable dimensions, do not identify new arrangements of variable combinations. A more comprehensive approach involves measuring multivariate distances across standardised variables (Williams et al. 2001, Elith et al. 2006, Roberts and Hamann 2012b, Mesgaran et al. 2014) or principal components (Broennimann et al. 2012, Eiserhardt et al. 2013; Box 4) or measurements to multivariate convex hulls around data clouds (Cornwell et al. 2006).

There are limited examples of cross-validations implemented with data splits directly in predictor space (Supplementary material Appendix 1 Table A1.2) and most are a byproduct of spatial data splitting (Fløjgaard et al. 2009, Roberts and Hamann 2012a, Wenger and Olden 2012). While Newbold et al. (2015) and Stephens et al. (2016) used biome delineations to block, these are also inferred extrapolations based on predefined spatial groups. Many assessments of model extrapolation fall under tests of the 'transferability' or 'generalisability' of a specific habitat model (Thomas and Bovee 1993, Leftwich et al. 1997, Schröder and Richter 2000, Chee and Elith 2012, Schibalski et al. 2014). While these studies evaluate model performances, they seldom quantify extrapolation requirements or analyse links between predictive performance decline and dissimilarity between training and prediction data. To date, while some ecological studies have linked decreases in predictive accuracy to measures of data dissimilarity (Capinha et al. 2012), only few have attempted to systematically quantify such patterns (Thuiller et al. 2004, Heikkinen et al. 2012, Roberts and Hamann 2012a, Bahn and McGill 2013), all of which expectedly found decreased prediction accuracy with increased dissimilarity between training and testing data. In these comparative studies, extrapolation was always a byproduct of spatial blocking. Of course, such validations assume that assessments of transferability in space to different predictor space can mimic assessments of transferability in time to different predictor space (Blois et al. 2013), but see (Schröder and Richter 2000).

How should blocks in predictor space be constructed?

A key challenge to blocking in predictor space for cross-validation is to decide how folds should be defined to inform the predictive objectives of the model. The intuitive approach may be to measure the dissimilarity between training and prediction data, then define blocks in such a way that the extrapolation requirements within the cross-validation are as similar as possible in magnitude and direction to those

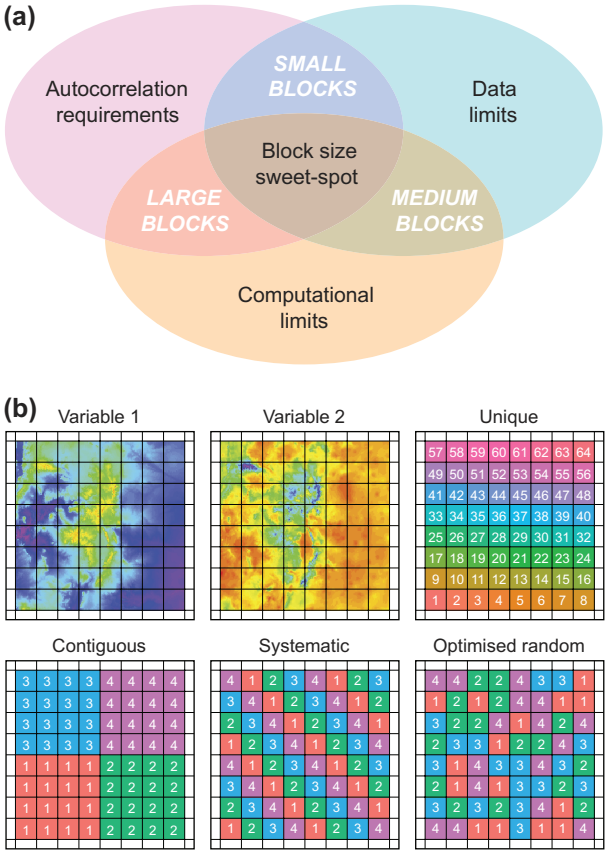


Figure 2. Approaches for choosing appropriate block sizes to minimise extrapolation. (a) The tradeoffs in block size selection between addressing residual autocorrelation requirements and working within the data and computational limits. (b) Sample fold assignments of hypothetical spatial blocks for cross-validation, which result in different levels of representation of predictor space based on combinations of two environmental variables (Variable 1 and Variable 2). When feasible, all blocks may be assigned to their own fold for cross-validation (Unique). When fewer folds than blocks are used, grouping contiguous blocks together will result in high dissimilarity between folds (Contiguous), except in very homogeneous environments. Both systematic assignment of folds (Systematic) and repeated random assignment with the final assignment based on minimum dissimilarity (Optimised random) can ensure lower dissimilarity between folds. While this figure shows a spatial example, similar approaches could be used for other correlation structures.

for the predictions. In *k*-fold approaches, where every fold is used exactly once for testing, this becomes a zero-sum game where the more one fold resembles the objective extrapolation, the more the others do not! In these cases, tests of predictions to particular folds may be more informative in the context of extrapolation than the overall error estimate across all folds.

Alternatively, cross-validations could be run several times with spatial or other structured blocks defined in a variety of sizes and/or orientations. This approach produces a range of validation statistics from the cross-validations rather than just a single value (Fig. 3a). From this range, it may be possible to define a limit, either in spatial block size or variable dissimilarity, at which the model no longer produces useful predictions. Such extrapolation limits, or 'forecast horizons', are common tools in economics (Ohlson and Zhang 1999), and meteorology (Foley et al. 2012) but have also recently been considered in ecology (Petchey et al. 2015; Fig. 3b). While this approach does not require any ad hoc calculation of dissimilarity, it is more computationally intensive in that dissimilarity measures and cross-validations must be undertaken for many blocking structures to determine the range of model performance.

To our knowledge, there are no examples in the literature of cross-validations using blocks purely defined in predictor space. In Box 4, we offer a species distribution modelling case study for North American Douglas-fir, in which we compare cross-validations based on random splitting, spatial blocking, and environmental blocking. Our results demonstrate that blocking purely in environment will decrease perceived model accuracy in cross-validations, but that estimates may remain optimistic if underlying correlation structures are not addressed.

Guidance: how to block

In this section, we suggest a workflow for cross-validation to clarify when and how to implement different data splitting strategies. The focus of this workflow is not on providing a fixed recipe for blocking, but rather on highlighting the questions a researcher should ask in this context. The exact answers to these questions are necessarily dependent on modelling objectives, data structures, computational capabilities, as well as the desire for conservatism in assessments of model forecast errors in the context of their results. We have discussed these implications and tradeoffs above.

Step 1. Assess dependence structures in the data

Determine the dependence structure in the raw data (temporal/spatial/phylogenetic autocorrelation using autocorrelation plots, variograms, or correlograms; quantify variance contribution in nested data using intercept-only mixed effect models). This serves as rough guidance on the scale of blocking (at least as many units as the range of autocorrelation; at least at the most variable hierarchical level). It should be emphasised here that, while modellers are most often concerned with autocorrelation in model residuals, dependence structures in this step are assessed on *raw data*, as this is where overfitting of predictor variables may occur.

Step 2. Determine prediction objectives

Will the model predict into new dependence structures (spaces, times, groups, etc.), or into new predictor space, or both? While extrapolation in predictor space, time, and geographic space may be straightforward to quantify (Box 1, 4), changes in hierarchical structure may necessitate more deliberation. For example, while some determinations of what constitutes a new 'group' of individuals may be obvious, others may be more nuanced (e.g. herds with non-overlapping ranges vs. individuals in the same herd that move largely independently, Box 2).

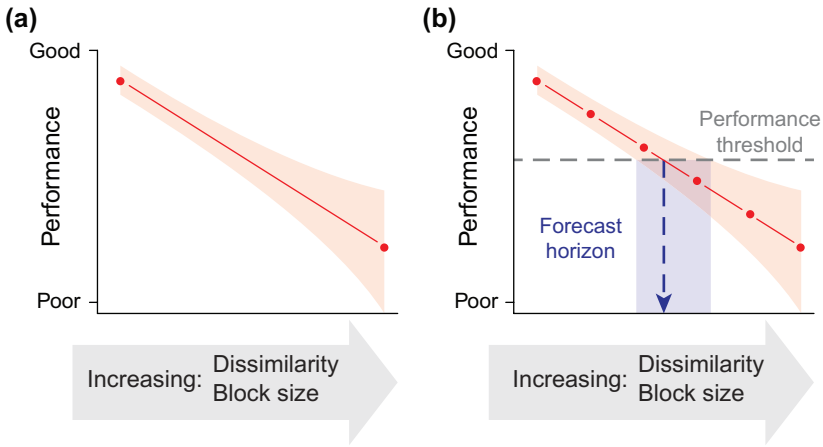


Figure 3. (a) Conceptual figure demonstrating the expected relationship between extrapolation requirements (either from dissimilarity between model training and prediction data or from blocking in cross-validation) and model accuracy, where accuracy generally decreases while dissimilarity increases. (b) Conceptual figure of a 'forecast horizon' (blue dashed line), based on repeated cross-validation with differently sized blocks, resulting from the drawing of a model performance threshold (dashed grey line). Because extrapolation requirements may vary at a given prediction performance threshold (e.g. across folds), the forecast horizon may include a range of values (blue shading). Extrapolations beyond the horizon would be considered too unreliable to be useful (adapted from Petchey et al. 2015).

Step 3. Block according to objectives and structure

When predictions will be made into new dependence structures, blocks should be drawn so that similar structural conditions are grouped together (e.g. spatial blocks when predicting to new sites; time-slice blocks of similar duration as the one predicting to; herds as blocks when predicting to a new herd; clades defined at the same phylogenetic branching depth as the clade to predict to; Box 1, Box 2, Box 3 for examples). Blocks can also be designed or arranged (or fold assignments or cross-validation methods can be chosen) to either minimise extrapolation or to emulate the extrapolation required between the training and prediction data (Box 4).

Step 4. Perform the cross-validation

Cross-validations may be performed for model comparison (and thus selection), error estimation, or both. There can be many blocks within each fold of a cross-validation. For example, if the block size of a spatial data set is 100×100 km, then the entire study region of Canada can be checkerboarded and a random set of blocks is assigned to each fold. Or, if within a herd, subgroups of ungulates with similar movement exist, these may serve as blocks and be assigned to fold across herds.

Step 5. Make 'final' predictions

The analyst essentially has two choices of how to determine a 'final model' or make 'final predictions'. Both have distinct advantages and disadvantages:

1) All the available training data can be used to fit a new model with which to make a single set of final predictions (Kuhn and Johnson 2013). As the error estimates from such a model are invalid, error estimates derived from the blocked cross-validation methods should be used. This approach favours final prediction quality over perfect accuracy of error estimates. It has the advantage of using all the data and thus likely being the best predictor, particularly for smaller datasets. It has the disadvantage that the error estimates from the cross-validation no longer apply perfectly to the predictions, as they were made with slightly different models. However it would be safe to assume that the error estimates are conservative (i.e. the final model should perform better), so this may not be a major disadvantage.
2) All the individual models from the cross-validation can be preserved and predictions from all models can be combined (Hastie et al. 2009). For example, in a *k*-fold approach, *k* different models are fitted, each describing a slightly different combination of training data. Predictions on the new data can be made with each of the *k* models, then averaged. This approach has the advantage of preserving the direct relationship between the models and the error estimates (i.e. the 'final models' are exactly those evaluated) as well as offering a variance for each prediction in the training data. On the downside, predictions are always made by models fitted with incomplete training data, compromising the sufficiency principle (i.e. that all possible information has been gleaned from the data) in the same way bagging does.

Challenges and limitations

While block cross-validation may helpful in situations with non-independent data, there are several instances in which spatial, temporal, phylogenetic, or blocking in predictor space may not be fruitful. This section aims at raising awareness of these problems and of the general limitations that prediction may face.

When data are scarce, cross-validation approaches that require models to be trained with further subset data may not be feasible. Similarly, even when data are numerous but only cover small spatial or temporal ranges, achieving independence between training and test data by blocking may not be possible. For example, if spatial autocorrelation persists at distances larger than half the spatial extent of the data, achieving independence in folds will be impossible no matter how spatial blocks are structured. This may also be the case for animal telemetry data when individuals move as a unified group. In such cases, no plethora of data records from within the same group will accommodate effective cross-validation for predictions to new independent individuals. This is more likely to occur within opportunistically collected data, than in data collected in systematic surveys.

Irregular sampling may lead to data clusters in space, time or along other correlation structures, which may lead to difficulties in defining effective regular blocks (Fig. 4a). In such cases, the models fit on training data may encounter highly variable sample sizes and prevalence rates, resulting in artificially large error estimates. One solution may be to use irregularly arranged but similarly sized blocks (Fithian et al. 2015; Fig. 4b) or irregularly shaped blocks (Lieske and Bender 2011; Fig. 4c).

Similarly, even when sampling coverage is unbiased and regular, in presence–absence data, prevalence of occurrences may be highly unbalanced (Fig. 4d), leading to blocks entirely lacking either presences or absences (e.g. withholding the centre square in Fig. 4d for validation). Unbalanced mean values of the response can also make cross-validation problematic if, for example, one tries to validate predictions using a block with only absence locations (Fig. 4d). While this may primarily be a presence–absence design problem, similar arrangements may appear in continuous response data. For example, in analyses with linear link functions, unequal means only affect estimates of the intercept, but for non-linear link functions (such as in count or presence–absence data) it affects all estimates.

Possible solutions for selecting evaluation blocks when data sampling is irregular or unbalanced could be: 1) non-gridded but consistently shaped blocks (e.g. pie-slices, Fig. 4e), 2) stratified blocks with similar mean/prevalence (Elith et al. 2008, who suggest a blocking strategy with equal number of occupied sites per block, Bahn and McGill 2013), 3) grouped sets of blocks (e.g. checkerboards; Box 1, 4) to ensure coverage of both presence–absence, and 4) buffered leave-one-out approaches (Bahn 2009, Telford and Birks 2009, Le Rest et al. 2014; Box 1, 4). It should be noted, however, when using non-regular spatial block shapes and arrangements, blocks may address autocorrelation inconsistently.

Last, in new predictor space, we might also encounter changes in relationships among covariates (changing correlation structures) or among species interactions themselves (Fielding and Bell 1997). This can be particularly problematic when predicting over larger time scales, enabling evolutionary changes to violate assumptions of niche conservatism (Maguire et al. 2015). Both situations, potentially undetectable by the modeller, can result in a loss of predictive power (Austin 2002) and are not, unfortunately, addressed through blocking in cross-validation.

Final thoughts

In this review and synthesis, we have discussed the role of block cross-validation for better estimating prediction errors. It addresses prediction optimism, arising from non-independent hold-out or from overfitting data dependence with covariates. We did not, however, attempt to address the effect of this overfitting on parameter estimation, or on model selection. These topics would benefit from further exploration.

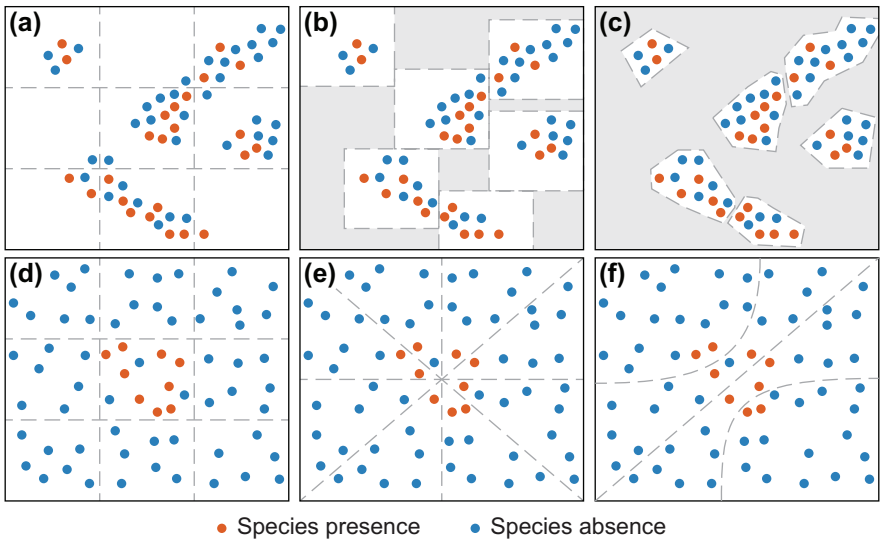


Figure 4. Conceptual illustrations of challenges for block cross-validation in space. Data may be (a) highly unbalanced in distribution of samples, which can be addressed through (b) irregular spacing of blocks of consistent size and shape, or (c) irregularly shaped blocks. Data may also be (d) highly unbalanced in prevalence (number of presences versus absences), which can be addressed through (e–f) non-gridded blocks and/or irregularly shaped blocks.

Statistical models in ecology are used not just to describe the present state of natural systems, but also to predict their change or development over time. Such models are fairly simple to create and have thus become ubiquitous in all areas of ecological research. To determine whether these statistical simplifications of ecological systems are useful, we need effective model validation procedures that produce reliable error estimates. Unfortunately, many popular evaluation and cross-validation approaches may result in erroneous and misleading assessments of model performance, either due to known and detectable issues of non-independence (e.g. residual autocorrelation) or due to more clandestine issues (e.g. structural parameterisation via covariates).

While developing a single-best approach to model validation applicable to all situations is impossible, informed choices may be guided by simple tests of dependence structures and extrapolation demands. Parametric measures of model fit or cross-validations with random data splits only provide reliable error estimates for model predictions in very specific cases where critical assumptions of independence and non-extrapolation are met. Such situations are rare in ecology and, as our simulations and case studies illustrate, can be very difficult to identify ad hoc.

In cases where the assumption of independence is compromised or where model extrapolation is likely, cross-validations with non-random blocks, carefully chosen in light of modelling objectives, can offer more reliable error estimates. The price of slightly conservative model validation (e.g. using a blocked approach when not necessary) is small compared to the unwarranted confidence in model predictions one might have with random cross-validation (or with no cross-validation at all). By overestimating predictive confidence, ecological modellers fail to adequately incorporate uncertainty into conservation and management decision-making and, more critically, sacrifice scientific credibility when a high proportion of such prognoses end up being wrong.

Data deposition

Data available from the Dryad Digital Repository: <http://dx.doi.org/10.5061/dryad.737gk> (Roberts et al. 2017).

Acknowledgements – We thank the German Science Foundation (DFG) for funding the workshop 'Model averaging in Ecology', held in Freiburg 2–6 March 2015 (DO 786/9-1), where the ideas included in this manuscript were developed. We thank Lara Budic for consultation and R code for phylogenetic trees and are grateful to all those who made their data available.

Funding – DRR is supported by the Alexander von Humboldt Foundation through the German Federal Ministry of Education and Research. BS is supported by the German Science Foundation (grant no. SCHR1000/6-2). CFD acknowledges additional funding by the DFG (DO786/10-1). DIW and JE are supported by Australian Research Council Future Fellowships (grant no. FT120100501 and FT0991640). GGA is the recipient of a Discovery Early Career Research Award from the Australian Research Council (project DE160100904). The work of JJLM was supported by the Australian Research Council Discovery Project DP160101003. Collection of data used to build elk resource selection models (Box 2) was funded by the Alberta Conservation Association (ACA – Grant Eligible Conservation Fund; grants to SC and MSB), the Natural Sciences and Engineering Research Council of Canada (NSERC CRD; grants to MSB and postdoctoral fellowship to SC), and Shell Canada limited. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions – All authors conceived the idea for this study. DRR, FH, CFD, SC and VB designed the study, and DRR, SC and VB carried out the simulations and analyses. The initial draft was written by DRR. All authors contributed comments and improvements to the manuscript.

References

Amesbury, M. J. et al. 2013. Statistical testing of a new testate amoeba-based transfer function for water-table depth reconstruction on ombrotrophic peatlands in northeastern Canada and Maine, United States. – J. Quat. Sci. 28: 27–39.

Anderson, R. P. 2013. A framework for using niche models to estimate impacts of climate change on species distributions. – Ann. N. Y. Acad. Sci. 1297: 8–28.

Araújo, M. B. et al. 2005. Validation of species-climate impact models under climate change. – Global Change Biol. 11: 1504–1513.

Arlot, S. and Celisse, A. 2010. A survey of cross-validation procedures for model selection. – Stat. Surv. 4: 40–79.

Austin, M. P. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. – Ecol. Modell. 157: 101–118.

Bahn, V. 2009. A new method for evaluating species distribution models. – 94th Ecol. Soc. Am. Ann. Meeting.

Bahn, V. and McGill, B. J. 2007. Can niche-based distribution models outperform spatial interpolation? – Global Ecol. Biogeogr. 16: 733–742.

Bahn, V. and McGill, B. J. 2013. Testing the predictive performance of distribution models. – Oikos 122: 321–331.

Bergmeir, C. and Benitez, J. M. 2012. On the use of cross-validation for time series predictor evaluation. – Inf. Sci. 191: 192–213.

Bjørnstad, O. N. and Grenfell, B. T. 2001. Noisy clockwork: time series analysis of population fluctuations in animals. – Science 293: 638–643.

Blois, J. L. et al. 2013. Space can substitute for time in predicting climate-change effects on biodiversity. – Proc. Natl Acad. Sci. USA 110: 9374–9379.

Boyce, M. S. et al. 2002. Evaluating resource selection functions. – Ecol. Modell. 157: 281–300.

Breiman, L. 2001. Random forests. – Machine Learning 45: 5–32.

Breslow, N. E. and Clayton, D. G. 1993. Approximate inference in generalized linear mixed models. – J. Am. Stat. Ass. 88: 9–25.

Brockwell, P. J. and Davis, R. A. 1996. Introduction to time series and forecasting. – Springer.

Broennimann, O. et al. 2012. Measuring ecological niche overlap from occurrence and spatial environmental data. – Global Ecol. Biogeogr. 21: 481–497.

Bulluck, L. et al. 2006. Spatial and temporal variations in species occurrence rate affect the accuracy of occurrence models. – Global Ecol. Biogeogr. 15: 27–38.

Burman, P. et al. 1994. A cross-validatory method for dependent data. – Biometrika 81: 351–358.

Buston, P. M. and Elith, J. 2011. Determinants of reproductive success in dominant pairs of clownfish: a boosted regression tree analysis. – J. Anim. Ecol. 80: 528–538.

Capinha, C. et al. 2012. Predicting the impact of climate change on the invasive decapods of the Iberian inland waters: an assessment of reliability. – Biol. Invas. 14: 1737–1751.
Chee, Y. E. and Elith, J. 2012. Spatial data for modelling and management of freshwater ecosystems. – Int. J. Geogr. Inf. Sci. 26: 2123–2140.
Coe, P. K. et al. 2011. Validation of elk resource selection models with spatially independent data. – J. Wildl. Manage. 75: 159–170.
Cornwell, W. K. et al. 2006. A trait-based test for habitat filtering: convex hull volume. – Ecology 87: 1465–1471.
Cressie, N. 1993. Statistics for spatial data. – Wiley.
Dormann, C. F. et al. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. – Ecography 30: 609–628.
Eiserhardt, W. L. et al. 2013. Dispersal and niche evolution jointly shape the geographic turnover of phylogenetic clades across continents. – Sci. Rep. UK 3: 1164.
Elith, J. and Graham, C. H. 2009. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. – Ecography 32: 66–77.
Elith, J. et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. – Ecography 29: 129–151.
Elith, J. et al. 2008. A working guide to boosted regression trees. – J. Anim. Ecol. 77: 802–813.
Elith, J. et al. 2010. The art of modelling range-shifting species. – Meth. Ecol Evol 1: 330–342.
Felsenstein, J. 1985. Phylogenies and the comparative method. – Am. Nat. 125: 1–15.
Fielding, A. H. and Bell, J. F. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. – Environ. Conserv. 24: 38–49.
Fithian, W. et al. 2015. Bias correction in species distribution models: pooling survey and collection data for multiple species. – Meth. Ecol. Evol. 6: 424–438.
Fløjgaard, C. et al. 2009. Ice age distributions of European small mammals: insights from species distribution modelling. – J. Biogeogr. 36: 1152–1163.
Foley, A. M. et al. 2012. Current methods and advances in forecasting of wind power generation. – Renew. Energ. 37: 1–8.
Fotheringham, A. S. et al. 2002. Geographically weighted regression: the analysis of spatially varying relationships. – Wiley.
Grafen, A. 1989. The phylogenetic regression. – Phil. Trans. R. Soc. B 326: 119–157.
Harvey, P. H. and Pagel, M. D. 1991. The comparative method in evolutionary biology. – Oxford Univ. Press.
Hastie, T. et al. 2009. The elements of statistical learning: data mining, inference and prediction. – Springer.
Hawkins, D. M. 2004. The problem of overfitting. – J. Chem. Inf. Comp. Sci. 44: 1–12.
Heikkinen, R. K. et al. 2012. Does the interpolation accuracy of species distribution models come at the expense of transferability? – Ecography 35: 276–288.
Ives, A. R. and Zhu, J. 2006. Statistics for correlated data: phylogenies, space and time. – Ecol. Appl. 16: 20–32.
Kennard, R. W. and Stone, L. A. 1969. Computer aided design of experiments. – Technometrics 11: 137–148.
Koenig, W. D. 1999. Spatial autocorrelation of ecological phenomena. – Trends Ecol. Evol. 14: 22–26.
Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. – 14th Int. Joint Conf. on Artifical Intelligence. Morgan Kaufmann Publ., San Francisco, CA, USA, pp. 1137–1143.
Koper, N. and Manseau, M. 2009. Generalized estimating equations and generalized linear mixed-effects models for modelling resource selection. – J. Appl. Ecol. 46: 590–599.
Kuhn, I. 2007. Incorporating spatial autocorrelation may invert observed patterns. – Divers. Distrib. 13: 66–69.
Kuhn, M. and Johnson, K. 2013. Applied predictive modeling. – Springer.
Larimore, W. E. and Mehra, R. K. 1985. The problem of overfitting data. – Byte 10: 167–178.
Le Rest, K. et al. 2014. Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. – Global Ecol. Biogeogr. 23: 811–820.
Leftwich, K. N. et al. 1997. Factors influencing behavior and transferability of habitat models for a benthic stream fish. – Trans. Am. Fish. Soc. 126: 725–734.
Legendre, P. 1993. Spatial autocorrelation – trouble or new paradigm. – Ecology 74: 1659–1673.
Legendre, P. and Fortin, M. J. 1989. Spatial pattern and ecological analysis. – Vegetatio 80: 107–138.
Lele, S. R. et al. 2013. Selection, use, choice and occupancy: clarifying concepts in resource selection studies. – J. Anim. Ecol. 82: 1183–1191.
Lieske, D. J. and Bender, D. J. 2011. A robust test of spatial predictive models: geographic cross-validation. – J. Environ. Inf. 17: 91–101.
Lundberg, P. et al. 2000. Population variability in space and time. – Trends Ecol. Evol. 15: 460–464.
Mackey, B. G. and Lindenmayer, D. B. 2001. Towards a hierarchical framework for modelling the spatial distribution of animals. – J. Biogeogr. 28: 1147–1166.
Maguire, K. C. et al. 2015. Modeling species and community responses to past, present, and future episodes of climatic and ecological change. – Annu. Rev. Ecol. Evol. Syst. 46: 343–368.
Mankin, J. B. et al. 1977. The importance of validation in ecosystem analysis. – In: Innis, G. (ed.), New directions in the analysis of ecological systems. Part I. The Society for Computer Simulation, pp. 63–71.
Manly, B. F. J. et al. 2002. Resource selection by animals: statistical design and analysis for field studies. – Kluwer.
Mesgaran, M. B. et al. 2014. Here be dragons: a tool for quantifying novelty due to covariate range and correlation change when projecting species distribution models. – Divers. Distrib. 20: 1147–1159.
Miller, J. et al. 2007. Incorporating spatial dependence in predictive vegetation models. – Ecol. Modell. 202: 225–242.
Mosteller, F. and Tukey, J. W. 1977. Data analysis and regression: a second course in statistics. – Addison Wesley.
Newbold, T. et al. 2015. Global effects of land use on local terrestrial biodiversity. – Nature 520: 45–50.
Ohlson, J. A. and Zhang, X. J. 1999. On the theory of forecast horizon in equity valuation. – J. Accounting Res. 37: 437–449.
Olden, J. D. et al. 2002. Predictive models of fish species distributions: a note on proper validation and chance predictions. – Trans. Am. Fish. Soc. 131: 329–336.
Otis, D. L. and White, G. C. 1999. Autocorrelation of location estimates and the analysis of radiotracking data. – J. Wildl. Manage. 63: 1039–1044.
Pearson, R. G. 2006. Climate change and the migration capacity of species. – Trends Ecol. Evol. 21: 111–113.
Petchey, O. L. et al. 2015. The ecological forecast horizon, and examples of its uses and determinants. – Ecol. Lett. 18: 597–611.
Picard, R. R. and Cook, R. D. 1984. Cross-validation of regression-models. – J. Am. Stat. Ass. 79: 575–583.
Power, M. 1993. The predictive validation of ecological and environmental-models. – Ecol. Modell. 68: 33–50.
Racine, J. 2000. Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. – J. Econometrics 99: 39–61.

Radosavljevic, A. and Anderson, R. P. 2014. Making better MAXENT models of species distributions: complexity, overfitting and evaluation. – J. Biogeogr. 41: 629–643.
Reineking, B. and Schröder, B. 2003. Computer-intensive methods in the analysis of species–habitat relationships. – In: Breckling, B. et al. (eds), Gene Bits und Ökosysteme: Implikationen neuer Technologien für die ökologische Theorie. Peter Lang, pp. 165–182.
Revell, L. J. 2010. Phylogenetic signal and linear regression on species data. – Meth. Ecol. Evol. 1: 319–329.
Roberts, D. R. and Hamann, A. 2012a. Method selection for species distribution modelling: are temporally or spatially independent evaluations necessary? – Ecography 35: 792–802.
Roberts, D. R. and Hamann, A. 2012b. Predicting potential climate change impacts with bioclimate envelope models: a palaeoecological perspective. – Global Ecol. Biogeogr. 21: 121–133.
Roberts, D. R. et al. 2016. Data from: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. – Dryad Digital Repository, <http://dx.doi.org/10.5061/dryad.737gk>.
Rooney, S. M. et al. 1998. Autocorrelated data in telemetry studies: time to independence and the problem of behavioural effects. – Mamm. Rev. 28: 89–98.
Rue, H. et al. 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. – J. R. Stat. Soc. B 71: 319–392.
Rykiel, E. J. 1996. Testing ecological models: the meaning of validation. – Ecol. Modell. 90: 229–244.
Schibalski, A. et al. 2014. Climate change shifts environmental space and limits transferability of treeline models. – Ecography 37: 321–335.
Schröder, B. and Richter, O. 2000. Are habitat models transferable in space and time? – J. Nat. Conserv. 8: 195–205.
Shao, J. 1993. Linear-model selection by cross-validation. – J. Am. Stat. Ass. 88: 486–494.
Snee, R. D. 1977. Validation of regression-models – methods and examples. – Technometrics 19: 415–428.
Stephens, P. A. et al. 2016. Consistent response of bird populations to climate change on two continents. – Science 352: 84–87.
Stone, M. 1974. Cross-validatory choice and assessment of statistical predictions. – J. R. Stat. Soc. B 36: 111–147.
Sumpter, D. J. T. 2006. The principles of collective animal behaviour. – Phil. Trans. R. Soc. B 361: 5–22.
Telford, R. J. and Birks, H. J. B. 2009. Evaluation of transfer functions in spatially structured environments. – Quat. Sci. Rev. 28: 1309–1316.
Telford, R. J. et al. 2004. Biases in the estimation of transfer function prediction errors. – Paleoceanography 19: PA4014.
Thomas, J. A. and Bovee, K. D. 1993. Application and testing of a procedure to evaluate transferability of habitat suitability criteria. – River Res. Appl. 8: 285–294.
Thuiller, W. et al. 2004. Effects of restricting environmental range of data to project current and future species distributions. – Ecography 27: 165–172.
Trachsel, M. and Telford, R. J. 2016. Technical note: estimating unbiased transfer-function performances in spatially structured environments. – Climate Past 12: 1215–1223.
Veloz, S. D. 2009. Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. – J. Biogeogr. 36: 2290–2299.
Verbyla, D. L. and Litvaitis, J. A. 1989. Resampling methods for evaluating classification accuracy of wildlife habitat models. – Environ. Manage. 13: 783–787.
Wenger, S. J. and Olden, J. D. 2012. Assessing transferability of ecological models: an underappreciated aspect of statistical validation. – Meth. Ecol. Evol. 3: 260–267.
Wiens, T. S. et al. 2008. Three way k-fold cross-validation of resource selection functions. – Ecol. Modell. 212: 244–255.
Williams, J. W. and Jackson, S. T. 2007. Novel climates, no-analog communities and ecological surprises. – Front. Ecol. Environ. 5: 475–482.
Williams, J. W. et al. 2001. Dissimilarity analyses of late-Quaternary vegetation and climate in eastern North America. – Ecology 82: 3346–3362.
Wu, J. G. and David, J. L. 2002. A spatially explicit hierarchical approach to modeling complex ecological systems: theory and applications. – Ecol. Modell. 153: 7–26.
Zurell, D. et al. 2012. Predicting to new environments: tools for visualizing model behaviour and impacts on mapped distributions. – Divers. Distrib. 18: 628–634.

Supplementary material (Appendix ecog-02881 at <www.ecography.org/appendix/ecog-02881>). Appendix 1: supplementary tables. Appendix 2: spatial blocking (Box 1), extended methods and detailed results. Appendix 3: blocking by individual or group (Box 2), extended methods and detailed results. Appendix 4: blocking to address phylogenetic correlation (Box 3), extended methods and detailed results. Appendix 5: blocking for extrapolation (Box 4), extended methods and detailed results. Appendix 6: R scripts and applicable data for the simulations and case studies.