Supplementary material

Carsten F. Dormann, Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R. García Marquéz, Bernd Gruber, Bruno Lafourcade, Pedro J. Leitão, Tamara Münkemüller, Colin McClean, Patrick E. Osborne, Björn Reineking, Boris Schröder, Andrew K. Skidmore, Damaris Zurell & Sven Lautenbach: Collinearity: a review of methods to deal with it and a simulation study evaluating their performance – Ecography 000: 000–000

Table of content

Appendix 1.1: Method descriptions

Table A1: List of collinearity methods used. Specific settings for individual methods are given in the appendix.

No.	*group*	*name*	*variations*	*Predictor types*	*Comments*	
1	Best possible	Truth		categorical and continuous	Best r^2 possible, given the noise in the simulated data	
2		GLM with correct model structure		categorical and continuous		
3-5 (AICc), 6-8 (BIC)	Clustering	PCA-based clustering	centred, summarised, explained	continuous	AICc and BIC	stepwise backward
9-11 (AICc), 12-14 (BIC)		Cluster analysis: Hoeffding-Ward	centred, summarised, explained	continuous	AICc and BIC	stepwise backward
15-17 (AICc), 18-20 (BIC)		Cluster analysis: Spearman-average	centred, summarised, explained	continuous	AICc and BIC	stepwise backward
21-23 (AICc), 24-26 (BIC)		iVIF	centred, summarised, explained	??	AICc and BIC	stepwise backward
27-28	Cluster independent	Select07			AICc and BIC	stepwise backward
29-30		Select04		Categorical and continuous	AICc and BIC	stepwise backward
31-32		Sequential regression		Categorical and continuous	AICc and BIC	Stepwise backward
33-34	Latent variable regressions	Principal Component Regression		Continuous	AICc and BIC	Stepwise backward
35-36		Constrained Principal Component Analysis		Continuous	AICc and BIC	Optimised
37-38		Partial Least Squares		Continuous	AICc and BIC	
39-40		Penalised Partial Least Squares		Continuous	AICc and BIC	Optimised
41		Latent Root Regression		Continuous		
42-43		DR/Sliced Inverse Regression		Categorical and continuous	AICc and BIC	stepwise backward
44		OSCAR		Categorical and continuous		Optimised
45-46	Robust methods	GLM		Categorical and continuous	AICc and BIC	Stepwise backward
47		GAM		Categorical and continuous		Spline shrinkage
48		LASSO		Categorical and contin-		Optimised

No.	*group*	*name*	*variations*	*Predictor types*	*Comments*	
				uous		
49		ridge		Categorical and continuous		Optimised
50		randomForest		Categorical and continuous		
51		Support Vector Machines		Categorical and continuous		ν (“nu”)-regression
52		Boosted Regression Trees		Categorical and continuous		
53		Multiple Adaptive Regression Splines		Categorical and continuous		
54-55		Collinearity Weighted Regression		Categorical and continuous	AICc and BIC	Stepwise backward

Identifying clusters

Full name of method: PCA-based clustering
Abbreviation: ./.
Key references: Booth et al. (1994)
Examples of implementation in ecology: ??
Brief description: In a first step, a principal component analysis is carried out (on the standardised predictors, i.e. using the correlation matrix). The loadings of each variable on each component are extracted. Starting with the first principal component, all variables loading on the same component with more than a pre-defined threshold are assigned to the same cluster. Hence, cluster groupings are achieved in a forward fashion. Backward PCA clustering starts with the least important PC, but derives more clusters and has a lower validation performance.

Once clusters are identified, they have to be represented (our options "summarised", "explained" and "centred") and then submitted to a regression model (such as a GLM, but other model types are conceivable).
Software used: R, with code written by CFD
Settings: default loading threshold of 0.32 (equivalent to 10% explained variance: Tabachnick and Fidell 1989)
Specifics of data manipulations for modelling: ./.
Predictors: continuous only, normally/symmetric distributed (ideally multinormal)
Response: Any type accepted by subsequent analysis (usually GLM).

Full name of method: cluster analysis
Abbreviation: ./.
Key references: Evritt et al. (2001); Kaufman & Rousseeuw (2005)
Examples of implementation in ecology: Pillar (1999)
Brief description: There are many different ways to use cluster analysis to identify clusters. The partitioning of variables into clusters is based on their relative distances (according to Euclidean, Chi-square, Bray-Curtis, Spearman, Pearson, Hoeffding and many others distances) and on the partitioning algorithm (single, complete, average, Ward, k-means and others).

Once clusters are identified, they have to be represented (our options "summarised", "explained" and "centred") and then submitted to a regression model (such as a GLM, but other model types are conceivable).
Software used: R with libraries Hmisc, with wrapper code written by CFD, BL and GC
Settings: We explored only two settings: Hoeffding-Ward and Spearman-average linkage. Both are options offered by function varclus in Hmisc, with Hoeffing-Ward as default. In either case, a threshold has to be set, at which level of similarity a cluster is partitioned off. For Hoeffding-Ward we used a threshold of 0, and for Spearman-average a value of 0.49 (this being based on a Spearman correlation between predictors of 0.7).
Specifics of data manipulations for modelling: ./.
Predictors: continuous and categorical
Response: Any type accepted by subsequent analysis (usually GLM).

Full name of method: iterative variance inflation factor analysis
Abbreviation: iVIF
Key references: Booth et al. (1994)
Examples of implementation in ecology: -
Brief description: The method works, essentially, by comparing the VIF values of a set of response variables with and without an additional response variable. All the variables that show an increase of the VIF value above a certain threshold and the newly added variable are grouped into one cluster (proxy-set in the terms of Booth et al. 1994).

In a first phase, all variables are added stepwise. If the introduction of a variable leads to a VIF value higher then a first threshold the variable stored in the proxy set of the collinear variable. Otherwise, the variable is added to a new proxy set. Only one representative member of each proxy set is tested in the first phase. If more then one variable shows collinearity towards a newly introduced variable all variables are moved to the proxy set of the first collinear variable. In a second phase, all variables from the proxy set are added again to the set of currently tested variables. Increasing VIF values (tested by a second threshold) indicated that the affected proxy sets have to be combined.
The iterative formula should ensure that all variable combinations are tested. The method identifies different groups than a classification based on pair-wise VIF values because it also considers the VIF of groups of more than two variables.
Software used: R with libraries car for calculation of VIF, code written by SL
Settings: We explored two settings, the threshold during the first phase and the second threshold which is applied during the second phase during the reinvention of the variables. Booth et al. (1994) suggests a value of 1.5 for the first threshold but do not specify the second threshold. In general, the lower the thresholds, the more variables will be grouped in proxy sets. Settings of 1.5 and 5 seemed to be fine for most situations. The second threshold is not well defined from a theoretical basis.

It is highly recommended to perform the IVIF only to a subset of variables which participate at least in on paired correlation coefficient greater then 0.5 or 0.7. Especially in situations there the correlated variables do not appear next to each other in the data frame, this preprocessing steps leads to more reliable results.
Specifics of data manipulations for modelling: ./.
Predictors: continuous and categorical
Response: Any type accepted by subsequent analysis (usually GLM).

Cluster-independent methods

Full name of method: selection of uncorrelated variables
Abbreviation: select07 (select04)

Key references: Green (1979); Fielding & Haworth (1995)
Examples of implementation in ecology: Oppel et al. (2004); Pearce-Higgins & Yalden (2004); Hernandez et al. (2006)
Brief description: Select07 is one of the most commonly used methods of removing collinearity from the input variable set. It returns a subset of the original predictor variables, only moderately correlated. Selection is based on a pairwise correlation matrix of the input variable set. From any two predictor variables exhibiting a correlation greater than the 'threshold value' the less important predictor variable with respect to the response variable is removed. Importance is determined either through ecological reasoning or by the deviances of univariate regressions of the predictors against the response variable. Five parameters need to be specified when using Select07 -threshold, method, sequence, family and univar. Fielding and Haworth (1995) recommend to use the threshold value |r|=0.7, although other, more conservative, threshold values have been suggested in the literature (cf. Capen et al. 1986). The correlation matrix can be calculated using either the Pearson's product-moment coefficient (`method="pearson"`) in case of bivariate normal distributions, or the non-parametric Spearman's rank correlation coefficient (`method="spearman"`), the default, if no assumptions are made about the frequency distribution of the variables. sequence can be a vector specifying the order of importance of the predictor variables obtained through ecological reasoning. If sequence=NULL (the default), the sequence of predictor variables is determined by their univariate importance for the response. In this case, family specifies the error distribution of the response variable, defaults to Gaussian. Univariate importances are calculated either by GLMs (`univar="glm"`), the default, or by GAMs (`univar="gam"`).
Collinearity method type: pre-analysis clean-up, returning an orthogonalised data set (as does PCA)
Software used: R, with code written by CFD, TM and DZ
Settings: We chose Pearson correlation threshold of 0.7 and 0.4, the first being commonly used, the second more restrictive. A threshold of 1 is equivalent to retaining all variables, i.e. use a GLM.
Specifics of data manipulations for modelling: ./.
Predictors: continuous and categorical
Response: Any type accepted by subsequent analysis (usually GLM).

Full name of method: Sequential regression
Abbreviation: ./.
Alternative names: residual regression
Key references: Graham (2003); Hastie et al. (2009, p. 53)
Examples of implementation in ecology: Graham (2003); Dormann et al. (2008a)
Brief description: Sequential regression builds new, orthogonalised predictor variables by removing from each predictor the fraction of variation already explained by more important variables. The sequence of importance is of the original predictor variables determined either through ecological reasoning or by the deviances of univariate regressions on the response variable. The most important predictor variable is kept as the first predictor variable of the new predictor set. Subsequently, following the sequence of importance, for each original predictor variable its independent contribution to the response is calculated by regressing it against all more important predictors of the new predictor set and replacing it with the residuals from the regression. The residuals from this regression form the corresponding predictor variables in the new predictor set. They are orthogonal, i.e. no longer statistically collinear, but conditional. Thus, they cannot be interpreted without the more important predictors. Likewise, stepwise procedures for model simplification can not be applied to the new predictor variables. Three parameters need to be specified when using SeqReg: family, univar and sequence. `family` specifies the error distribution of the response variable, defaults to Gaussian. `sequence` can be a vector specifying a sequence of importance of the predictor variables based on ecological reasoning. If sequence=NULL (the default), the sequence of predictor variables is determined by their univariate importance for the response. In this case, two methods are available for calculating the univariate (=marginal) importance of the predictor variables for the response, GLM (univar="glm"), the default, and GAM (univar="gam").

A stepwise model simplification is very computer intensive, because after removing a variable, all variables of lower importance have to be re-calculated. The interpretation of variables changes from "there is a positive effect of precipitation" to "there is a precipitation effect additional to the contribution it already made through its correlation with temperature". Conceptually, sequential regression is related to semi-partial correlation analysis (Bortz 1993) and path analysis, where also variables can act through correlation with other variables (Grace 2006).

To make predictions with the results of a sequential regression requires the similar processing of the data set onto which to predict. Thus, the sequence of variables from the original regression is passed onto the new data set, and this will then in turn be sequentially regressed to yield a modified new data set of predictor variables. This is then used for prediction.
Collinearity method type: pre-analysis clean-up, returning an orthogonalised data set (as does PCA)
Software used: R, code written by CFD and JG
Settings: defaults
Specifics of data manipulations for modelling: ./.
Predictors: continuous and categorical
Response: Any distribution accepted by a GLM.

Latent variable modelling

Full name of method: Principal Component Regression
Abbreviation: PCR
Key references: Jackson (1991); Rawlings (1988)
Examples of implementation in ecology: Riffell & Gutzwiller (2009)

Brief description: Principal Component Regression is an interconnection of Principal Component Analysis (PCA) and multiple linear regression.

First, the principal components are calculated. We formalize this concept. The goal is: Find a set of weights *w* in order to create a linear combination of the columns of X, i.e. $\xi = Xw$, such that its variance is maximum. This goal can be attained by maximization with respect to *w* of $w^T X^T Xw$ under the constraint that weight vector *w* is normalized. This problem leads to an eigenanalysis solution. Here, *w* is the first eigenvector of $X^T X$.

W is the matrix formed by all eigenvectors of $X^T X$. Thus *W* is orthonormal. Furthermore, all components ξ_i are mutually orthogonal. Ξ is the matrix formed by all these principal components.

Second, the most important principal components are used as regressors to fit the responses *y* by a multiple linear regression. This means, that the key point is to find *m* eigenvectors that can be used to calculate principal components $\Xi_m = XW_m$ as an optimal set for regression $y = \Xi_m \ \beta_{PCA} + \varepsilon$. Then, the OLS estimator is the solution of the minimization problem $\min \ (y - \hat{y})^T (y - \hat{y})$. Therefore, OLS prediction for *y* is: $\hat{y}_{PCR}{}^m = \Xi_m \ (\Xi_m{}^T \Xi_m)^{-1} \Xi_m{}^T y$.

Software used: R with library stats, with wrapper code written by CFD

Settings: The variables should be scaled in function prcomp. One parameter (number of principal components) has to be specified when using PCR. In the wrapper code, the function step is used to select an optimal set of principal components by AIC.

Specifics of data manipulations for modelling: ./.

Predictors: continuous

Response: Any type accepted by subsequent analysis (usually GLM).

Full name of method: Partial Least Square regression

Abbreviation: PLS

Key references: Jackson (1991); Martens & Naes (1989)

Examples of implementation in ecology: Carrascal et al. (2009)

Brief description: In PCA the first principal components are appropriate variables to explain *X*. Nothing guarantees that these principal components are relevant for *y*. Partial Least Squares (PLS) regression attempts to set up a model using both the predictors *X* and the responses *y*. Therefore, PLS regression generalises PCA and combines it with a multiple linear regression. PLS regression is an iterative procedure.

We now formalize this concept. The goal is: Find a set of weights *w* in order to create a linear combination of the columns of X, i.e. $t = Xw$, such that their covariance to *y* is maximum.

This goal can be attained by maximization with respect to *w* of $w^T X^T yy^T Xw$ under the constraints that weight vector *w* is normalized and the matrix $T = (t_1, t_2, \ldots)$ is orthogonal. This problem leads to an eigenanalysis solution. The vector *w* is the first eigenvector of $X^T yy^T X$. Furthermore, a deflation of *X* ensures that a given component t_i is orthogonal to all others. That is, when the first vector t_1 is found, its projection onto *X* is subtracted from *X*. Then, the procedure is re-iterated until *X* becomes a null matrix.

For regression the first *m* vectors of the so-called latent variables: $T = (t_1, t_2, \ldots, t_m)$ are used as an optimal set. Therefore, OLS prediction for *y* is: $\hat{y}_{PLS}{}^m = T_m \ (T_m{}^T T_m)^{-1} T_m{}^T y$.

Software used: R with libraries ppls and gpls, with wrapper code written by GC and CFD

Settings: One parameter (number of latent variables) has to be specified when using PLS.

The wrapper code calculates AIC for all fits and finds the most parsimonious model, i.e. the appropriate number of latent variables *m*.

Specifics of data manipulations for modelling: ./.

Predictors: continuous

Response: Gaussian or binomial

Full name of method: Penalized Partial Least Squares

Abbreviation: PPLS

Key references: Krämer et al. (2007)

Examples of implementation in ecology: ?

Brief description: PCR, PLS, LRR and CPCA have in common that their second step is a linear regression, i.e. they are linear approaches. Nonlinear regressions may be fitted by use of additive regression models where the linear predictor is substituted by a user specified sum of smooth functions.

Here penalized regression splines are a prevalently used tool. Splines are piecewise polynomial functions. Their links are called knots.

The main idea of Penalized PLS is to use such nonlinear splines transformation $Z = \Phi_B(X)$ as a preliminary step followed by the linear PLS approach. For performing this transformation, B-splines are used as a set of basis functions. If *X* has dimensions $n \times p$, then *Z* has dimensions $n \times pK$, where *K* depends on the order of splines and the number of knots.

One has to take into account that the OLS method of regression by minimization with respect to β of $(y - Z\beta)^T (y - Z\beta)$ possibly leads to overfitting because of the high-dimensional matrix of predictors *Z*. Here, the number of variables is generally larger than the available observations. In order to tackle this problem, the principle of penalization (e.g., see ridge regression) is adapted for PPLS. Therefore, the OLS method is substituted by minimization with respect to β of $(y - Z\beta)^T (y - Z\beta) + \beta^T P \beta$, where *P* is a penalty matrix.

In summary, penalized PLS means that PLS is performed on the transformed matrix *Z* and the PLS regression is performed with penalty matrix *P*.

Software used: R with library ppls, with wrapper codes written by GC and CFD
Settings: Three parameters (the number of latent variables, the number of knots, and a parameter for the penalty matrix) have to be specified when using PPLS. The wrapper code includes a helper function for the optimization of the number of knots and the parameter for the penalty matrix. Moreover, this code calculates AIC for all fits and finds the most parsimonious model, i.e. the appropriate number of latent variables *m*.
Specifics of data manipulations for modelling: ./.
Predictors: continuous
Response: Gaussian

Full name of method: Constrained Principle Component Analysis
Abbreviation: CPCA
Key references: Vigneau et al. (2002)
Examples of implementation in ecology: ?
Brief description: The main improvement of PLS versus PCR is that the variance of *X* is substituted by the covariance of *X* and *y*. This concept should confirm that the chosen latent variables are relevant not only for *X*, but also for *y*. Regression models through Constrained Principal Components Analysis similarly attempt to create latent variables using both the predictors *X* and the response *y* . In contrast to PLS, however, CPCA is a straightforward model, i.e. it is non-iterative. Furthermore, a tuning parameter α is introduced in CPCA.

We now formalize this concept. The goal is: Find a set of weights *l* in order to create a linear combination of the columns of *X*, i.e. $z = Xl$, such that it provides the best prediction of matrix $A = \left[\alpha\, y \,|\, (1-\alpha)X\right]$. Its OLS estimator is the solution of the minimization problem $\min\,(A-\hat{A})^T(A-\hat{A})$. This goal can be attained by maximization with respect to *U* of $trace(U^T V_X^{-1/2} V_{XA} V_{AX} V_X^{-1/2} U)$ under the constraint that matrix *U* is orthonormal, where $V_X = X^T X,\ \ V_{XA} = X^T A,\ \ V_{AX} = A^T X$. This problem leads to an eigenanalysis solution:
U is the matrix formed by the eigenvectors of $V_X^{-1/2} V_{XA} V_{AX} V_X^{-1/2}$.

For regression the first *m* eigenvectors $Z_m = XL_m = XV_X^{-1/2} U_m$ are used as an optimal set. Therefore, OLS prediction for *y* is: $\hat{y}_{CPCA}{}^m = Z_m\,(Z_m{}^T Z_m)^{-1} Z_m{}^T y$.
Software used: R code written by GC and wrapper code written by CFD
Settings: Two parameters (number of latent variables, tuning parameter) have to be specified when using CPCA. The wrapper code calculates AIC for all fits and finds the most parsimonious model, i.e. the appropriate number of latent variables *m*. Systematic evaluation of critical parameter value for the tuning parameter yields the rule of thumb: `alpha=0.1`
Specifics of data manipulations for modelling: ./.
Predictors: continuous
Response: Gaussian

Full name of method: Latent root regression
Abbreviation: LLR
Key references: Vigneau & Qannari (2002)
Examples of implementation in ecology: ?
Brief description: The main improvement of PLS versus PCR is that the variance of *X* is substituted by the covariance of *X* and *y*. This concept should confirm that the chosen latent variables are relevant not only for *X*, but also for *y*. Latent root regression similarly attempt to create latent variables using both the predictors *X* and the responses *y*.

For this purpose we define a matrix *B* as follows: $B = \left[y \,|\, \Xi\right]$, where Ξ is the matrix of principal components associated with *X* (see PCR). LRR is based on a discussion of the relevance (i.e. predictive value for *y*) of principal components associated with *B*. By means of a deflation procedure like in PLS, it is possible to stepwisely select latent variables. This iterative procedure adopts some matrix properties of *B* and Ξ and therefore circumvents a detailed investigation regarding predictive value.
Software used: R code written by GC, PJL, and CFD
Settings: One parameter (number of latent variables) has to be specified when using LRR. The function step is used to select an optimal number by AIC.
Specifics of data manipulations for modelling: ./.
Predictors: continuous
Response: Gaussian

Full name of method: Dimension reduction regression
Abbreviation: DR
Alternative names: Sliced inverse regression (SIR)
Key references: Cook & Weisberg (1999); Weisberg (2008)
Examples of implementation in ecology: ./.
Brief description: DR is a set of dimension reduction techniques in which the minimum underlying subspace of the data is to be estimated by kernel approaches. Different objective functions and kernel approaches are implemented (Weisberg 2008).
Software used: R, with the library dr and additional code by BR and BS
Settings: `slice.function=dr.slices.arc, nslices=8, chi2approx="wood", method="sir"`
Specifics of data manipulations for modelling: ./.
Predictors: continuous and categorical (only for some methods within dr)
Response: any distribution accepted by a GLM

Tolerant methods

Full name of method: Ridge regression
Abbreviation: ridge
Alternative names: ./.
Key references: Hoerl & Kennard (1970)
Examples of implementation in ecology: Reineking & Schröder (2006)
Brief description: Ridge regression is a penalized regression method – other examples are LASSO or OSCAR. Penalized regression methods aim at improving

ordinary least squares (OLS) estimates in regard to the accuracy of the model and the interpretability of the model. This is achieved by adding a penalization term to the model, which penalizes large coefficients. Model selection favours models with smaller coefficients if these can achieve a similar quality of the fit than models with larger coefficients. Ridge regression puts a penalty in form of the L_2-norm of the regression coefficients $\hat{\beta}_j$ to the OLS estimation:

$$L(\lambda_2, \beta) = \|y - X\beta\|^2 + \lambda_2 |\beta|^2$$

$$|\beta|^2 = \sum_{j=1}^{p} \beta_j^2$$

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|^2, \text{ subject to } \sum_{j=1}^{p} \beta_j^2 \leq t$$

Before model is fitted, the response is centred and the predictors are standardized. The choice of tuning parameters λ is typically achieved by cross-validation. The use of the L_2 norm leads to a shrinkage of the coefficients towards zero. This shrinkage leads to an additional estimation bias but on the other hand results in a smaller prediction error due to increased variance (Hastie et al. 2009).

The effective degrees of freedom for the ridge regression can be calculated by:

$$df(\lambda) = tr(X(X^T X + \lambda I)^{-1} X^T) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}$$

(Hastie et al. 2009), with d_j being the jth element of the diagonal matrix of the singular value decomposition of the centred input matrix X.

Software used: R, together with the library penalized
Settings: `minlambda1=0, maxlambda1=100, fold=10, standardize=TRUE`
Specifics of data manipulations for modelling: ./.
Predictors: continuous and categorical
Response: any distribution accepted by a GLM

Full name of method: LASSO
Abbreviation: ./.
Alternative names: ./.
Key references: Tibshirani (1996)
Examples of implementation in ecology: Reineking & Schröder (2006)
Brief description: LASSO is a penalized regression method (see ridge regression above). The penalization term added to the OLS model is the L_1 norm:

$$L(\lambda_1, \beta) = \|y - X\beta\|^2 + \lambda_1 |\beta|_1$$

$$|\beta|_1 = \sum_{j=1}^{p} |\beta_j|$$

which is equivalent to the optimization problem:

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|^2, \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq t$$

Before model is fitted, the response is centred and the predictors are standardized. The choice of tuning parameters is done typically done by cross-validation. The use of the L_1 norm leads in addition to the shrinkage of the model coefficients towards zero to a selection process between the explanatory variables. For the lasso the number of non-zero coefficients is a valid estimator for the degrees of freedom of the model (Zou and Hastie 2005).

Software used: R, with the library penalized
Settings: `minlambda1=1e-8, maxlambda1=100, fold=10, standardize=TRUE`
Specifics of data manipulations for modelling: ./.
Predictors: continuous and categorical
Response: any distribution accepted by a GLM

Full name of method: octagonal shrinkage and clustering algorithm for regression
Abbreviation: OSCAR
Alternative names: ./.
Key references: Bondell & Reich (2007)
Examples of implementation in ecology: -
Brief description: The OSCAR uses the the L_1-norm together with the pair-wise L_∞ -norm on $\hat{\beta}_j$ as penalization terms:

$$L(\lambda, \beta) = \|y - X\beta\|^2 + \lambda \left[|\beta| \;\; + c \sum_{j<k} \max\{|\beta_j|, |\beta_k|\} \right]$$

which becomes for ordered β_j ($|\beta_1| \leq |\beta_1| \leq \ldots |\beta_p|$)

$$L(\lambda, \beta) = \|y - X\beta\|^2 + \lambda \left[(c(j-1)+1) \sum_{j=1}^{p} |\beta_j| \right]$$

which is equivalent to $(1 - c)\, L_1 + c$ (pairwise L_∞) $\leq t$, and which is equivalent to the optimization problem:

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|^2, \text{ subject to}$$

$$(1-c)|\beta|_1 + c \sum_{j<k} \max\{|\beta_j|, |\beta_k|\} \leq t$$

Before the model is fitted, the response is centred and the predictors are standardized. The choice of tuning parameters is typically achieved by cross-validation. Since both terms have to be optimized in parallel and since the calculations involved are computational intense using the standard Matlab solver the model fitting procedure is time consuming. Upcoming developments might improve this (Bondell and Reich 2007). The use of the L_1-norm leads, in addition to the shrinkage of the model coefficients towards zero, to a selection process among the explanatory variables. By use of the pairwise L_∞ -norm the OSCAR encourages both sparsity and equality of coefficients (for standardized predictors). Sparsity leads to grouping of highly correlated predictors and performs thereby an additional clustering of the predictor variables. For the OSCAR estimate of the degrees of freedom for the OSCAR is the number of distinct nonzero values of $\{|\hat{\beta}_1|, \ldots, |\hat{\beta}_p|\}$ (Bondell and Reich 2007).

Software used: R, Matlab code from (Bondell and Reich 2007), code to call compiled Matlab executable and model selection by SL
Settings: defaults
Specifics of data manipulations for modelling: ./.
Predictors: continuous

Response: Gaussian, an extension towards the distributions accepted by a GLM is planed (Bondell and Reich 2007).

Full name of method: Boosted Regression Trees
Abbreviation: BRT
Alternative names: stochastic gradient boosting
Key references: Friedman (2000); Schapire (2002); Elith et al. (2008)
Examples of implementation in ecology: Leathwick et al. (2006a); Elith & Leathwick (2007)
Brief description: Boosted regression trees combine two algorithms: "boosting" is a method for developing multiple models and combining them; "regression trees" are single models that partition the predictor space into disjoint regions and predict a separate constant value in each of them (Friedman and Meulman 2003). The boosting algorithm calls the regression tree algorithm repeatedly, each time giving it a re-weighted version of the data that emphasizes the records that were misclassified in the last round. Finally the suite of trees is combined by weighted averaging (Schapire 2003). Statisticians have reinterpreted it as a method for developing a regression model in a forward stage-wise fashion, adding small modifications across the model space (via trees) to fit the data better (Hastie et al. 2009). The final model has numerous terms, each term being a regression tree. As boosting proceeds, the model complexity increases until eventually it over-fits the data. In the gradient boosted methods (Friedman 2002) the aim is to maximize the log-likelihood, and updates are based on its gradient. The number of trees in the boosted model is a natural measure of complexity, and is chosen by measuring prediction accuracy on independent data (cross-validation). This identifies the most complex model that still predicts well, and is based on the trade-off between training error and generalization error. The two main parameters to be set are the shrinkage parameter (learning rate), which controls the amount of re-weighting at each step, and the size of each tree – one partition (an additive model) or two or more splits. BRT is implemented in gbm (see below) for several response types, including binomial families.
Software used: R with library gbm; extra code written by JE and John Leathwick
Settings: defaults, but `learning.rate = 0.01, tree.complexity = 5`
Specifics of data manipulations for modelling: ./.
Predictors: continuous and categorical
Response: binary/multinomial (classification) or continuous (regression; Gaussian or Poisson).

Full name of method: randomForest
Abbreviation: ./.
Alternative names: ./.
Key references: Breiman (2001)
Examples of implementation in ecology: Cutler et al. (2007)
Brief description: randomForest is a machine-learning algorithm building on repeated construction of classification and regression trees based on randomly drawn variables and samples (Breiman 2001). Each tree is then validated on the cases withheld from fitting (out-of-bag validation). M trees are "grown" and all are used for prediction, weighted by their validation performance. randomForest is fast, robust and easy to employ. Three parameters have to be specified when using randomForest: mtry ("Number of variables randomly sampled as candidates at each split. Default values are different for classification (sqrt(p), where p is number of variables) and regression (p/3)", citing the function's help-page in R. However, for the simulated data used here, more tries yielded better predictions, usually levelling off at p/2, but still being best at p.); ntree (number of trees to grow; defaults to 500, which also pre-trials confirmed to be a good value); and nodesize (minimum size of terminal nodes; the larger this value, the shorter the trees; defaults of 1 and 5 (for classification and regression, respectively) did not yield the best predictions in our pre-trials, rather, 1, 2, 4 and 7 performed best).
Collinearity method type: robust (i.e. does not make specific adjustments for collinearity of predictors)
Software used: R, with library randomForest
Settings: defaults, but mtry=20 and ntree=1000 (We ran several trials on the same data before the study, with "test same" as test data set, to find the best settings for mtry and ntree. The outcome wea not particularly clear for ntree, but larger values of mtry led to better fits on the test data. The chosen values for mtry are rather high and may contribute to the overfitting we detect in the study for altered collinearity structure.)
Specifics of data manipulations for modelling: ./.
Predictors: continuous and categorical
Response: normal (regression), categorical (classification)

Full name of method: Support Vector Machines
Abbreviation: SVM
Alternative names: ./.
Key references: Vapnik (1996); Burgess (1998); Hastie et al. (2009); Steinwart & Christmann (2008)
Examples of implementation in ecology: Guo et al. (2005); Ribeiro & Torgo (2008)
Brief description: Support Vector Machines were initially conceived as multi-dimensional classifiers. They separate samples in n-dimensional space by finding a "hyperplane" that separates the different classes in such a way as to maximise the distance between the *nearest* points of different classes. To do so, the model starts with a linear separation and then uses kernels to increase the complexity of the separator. The points that form the border of a class are called "support vectors", hence the name. One key feature in this method is that any point beyond the class boundary is irrelevant for the calculations. Since SVM find the optimum separation planes iteratively, at some point during the model building every data point has been used.
In *support vector regression*, a distance-dependent loss function (Gaussian, Huber, Laplace or other) replaces the separation line and weights the contribution of each data point. The ε-insensitive loss function actually defines a region of size ε where data points are ignored, corresponding to the "already perfectly classified data" in the classification model. In ν ("nu")-regression, the

parameter ν controls the “data-ignoring width” ε and can be optimised (Schölkopf and Smola 2000).
Collinearity method type: robust (i.e. does not make specific adjustments for collinearity of predictors)
Software used: R, with library e1071
Settings: defaults but with `type=”nu-regression”`. Default settings include scaling of the variables and the use of the radial basis kernel.
Specifics of data manipulations for modelling: ./.
Predictors: continuous and categorical
Response: normal (regression), categorical (classification)

Full name of method: Multivariate Adaptive Regression Splines
Abbreviation: MARS
Alternative names: earth (for MARS being a registered acronym)
Implementation in this study: single species models
Key references: Friedman (1991); Hastie et al. (2009)
Examples of implementation in ecology: (Moisen and Frescino 2002, Yen et al. 2004), (Leathwick et al. 2006b) (Elith and Leathwick 2007, Heinänen and Numers 2009)
Brief description: MARS is a hybrid between conventional regression and recursive partitioning methods. MARS uses piece-wise linear basis functions to define the modelled relationship. Basis functions are defined in pairs, using a knot to define inflection points, and coefficients to quantify the slopes of the non-zero sections. More than one knot (i.e. more than one pair of basis functions) can be specified for a predictor variable, allowing complex non-linear relationships to be fitted. When fitting a MARS model, knots are chosen in a forward stepwise procedure. Candidate knots can be placed at any position within the range of each predictor variable to define a pair of basis functions. At each step, the model selects the knot and its corresponding pair of basis functions that give the greatest decrease in the residual sum of squares. Knot selection proceeds until some maximum model size is reached, after which a backwards-pruning procedure is applied and those basis functions that contribute least to model fit are progressively removed. At this stage, a predictor variable can be dropped from the model completely if none of its basis functions contribute meaningfully to predictive performance. The sequence of models generated from this process is then evaluated using generalized cross-validation, and the model with the best predictive fit is selected. Interactions between variables can be fitted, but rather than fitting a global interaction between a pair of variables, these are specified for only part of the environmental range using basis functions. The current implementation of MARS in R uses least squares fitting appropriate for data with normally distributed errors.
Collinearity method type: robust (i.e. does not make specific adjustments for collinearity of predictors)
Software used: R, with library mda; additional code written by J. Leathwick and J. Elith
Settings: defaults
Specifics of data manipulations for modelling: ./.
Predictors: continuous and categorical
Response: Any distribution accepted by a GLM.

Full name of method: collinearity-weighted regression
Abbreviation: CWR
Alternative names: ./.
Key references: (this is a new methods, described here for the first time)
Examples of implementation in ecology: ./.
Brief description: Collinearity may be caused by some few data points (e.g. outliers). Reducing their importance in the regression will also yield parameter estimators less affected by collinearity. Following this line of argument, CWR calculates a vector of weights that reduces the influence of data points causing collinearity. These weights can be passed on to a GLM model. For calculation of weights, each data point is omitted in turn from the regression, and then the maximum condition index for the model is calculated. Maximum CI values are then taken to the power of log(N) (to give small changes less importance in small data sets than in large data sets) and assigned as preliminary weights to the data point omitted. Finally, these preliminary weights are divided by the mean of preliminary weights of all points and assigned as weights to be used in the further analysis.
Collinearity method type: incorporate collinearity (GLM with specific collinearity adjustment)
Software used: R, with code written by TM, BR and CFD (giveWeightsVif)
Settings: defaults
Specifics of data manipulations for modelling: ./.
Predictors: continuous and categorical
Response: any distribution accepted by a GLM

Appendix A2: Comparison of methods

Simulated data

Details on data simulation

Collinear data creation

All data sets consisted of 1000 data points and 21 explanatory variables (predictors). Predictors were grouped into four clusters of five variables, plus a single, uncorrelated variable, and collinearity was restricted to within clusters. To achieve collinearity, one predictor within a cluster was drawn from a uniform distribution between 0 and 1. The second variable in the cluster was simply the first with a random normal noise (mean = 0) added to it, with the standard deviation of the normal noise determined by a free parameter called "decay". The third variable of the cluster was analogously derived from the second, and so forth. The parameter "decay" hence determined how fast the collinearity decreases from the first to the fifth predictor in a set: high decay means low collinearity. The 21st predictor was always created as uncorrelated with all others. All X were then standardised.

Simulation analysis

For all methods dealing with collinearity before the analysis and hence creating a new data set (e.g. select07, clustering methods, some latent variable methods, see supplementary material for details of each method), we used a linear model with Gaussian errors, linear and quadratic terms, interactions and stepwise model simplification to analyse the data (from here on referred to as GLM). Many analyses used an information-theoretic threshold for selecting the final model (either for deciding on the number of components or for stepwise model selection; see Table C1 in supplementary material). The most common approach was to use (possibly small-sample size-corrected) Akaike's Information Criterion (AICc, Burnham and Anderson 2002), although the AIC has been criticised as being asymptotically biased (e.g. Link and Barker 2006). We therefore used both AICc and the Bayesian Information Criterion (BICc) for model selection, the latter resulting usually in more parsimonious models.

Several collinearity approaches required the specification of parameters (such as constraints, weights, penalties and so forth). These were optimised for each data set using only the training data. Selecting the best model through parameter tuning and model selection led to a considerable computational burden. The analysis of a single data set on a standard desktop computer took seven hours, or 28 000 hours for all data sets.

Latent variable regression models are commonly carried out without consideration of non-linear relationships or interactions between latent variables. Although this could easily be incorporated in the analysis, we have not found any study that actually did so. We did use linear and quadratic combinations of latent variables in their linear models, because a pre-analysis showed linear combinations to be highly inadequate. We did not, however, use interactions between latent variables. This may lead to a slight reduction in performance compared to more completely defined GLMs and the like.

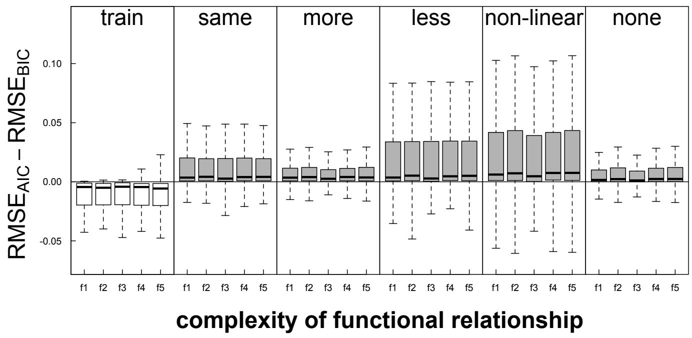
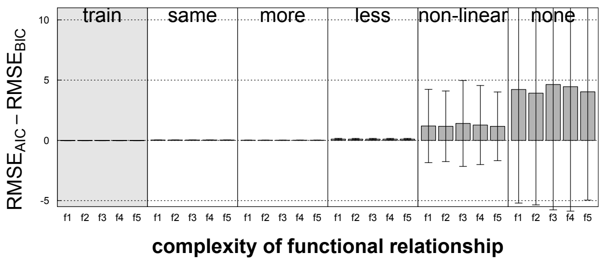


Fig. A1. Difference in Root Mean Square Error of AIC- and BIC-derived models, averaged across all 23 methods. Positive values indicate overfitting by AIC-derived models. Data points outside 1.5 x interquartile range omitted for clarity.

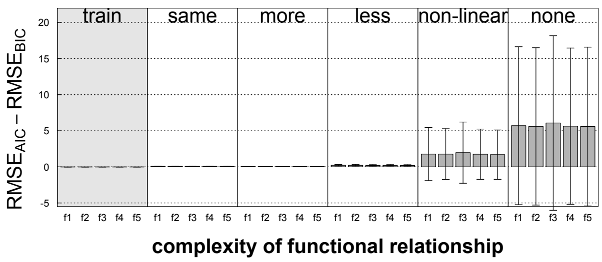
Pre-analysis: AIC vs. BIC

Across all methods, BIC-derived models were consistently more accurate in predicting test data (i.e. RMSE-difference between AIC and BIC is positive, indicating higher errors in the AIC-based predictions: Fig. A1). However, variation was huge and dependent on both the type of test data and the method. For some methods, the difference between AIC- and BIC-derived models was negligible (all clustering methods, PCR and DR), and for CPCA, PLS and PPLS AIC-derived models were slightly better than BIC-derived. For all other methods, BIC yielded higher correlation coefficients between predictions and test data. The difference between training fit and test fit was particularly large for GLM, seqreg and CWR when using AIC, indicating substantial overfitting (Fig. A2). The BIC-versions of these approaches, in contrast, show virtually no overfitting.

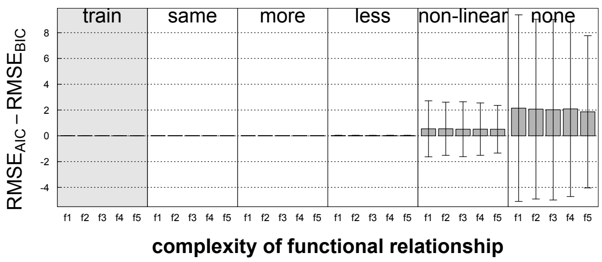
GLM



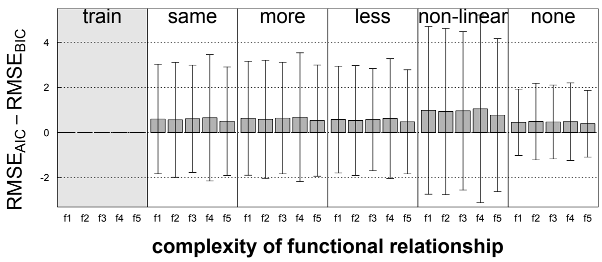
CWR



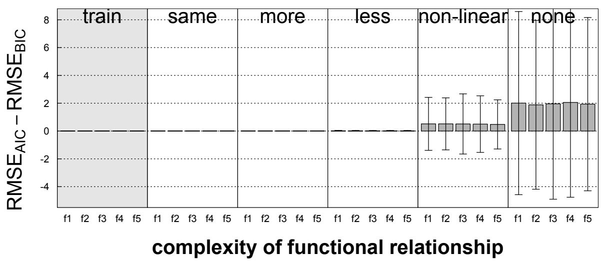
clust:PCA:cent



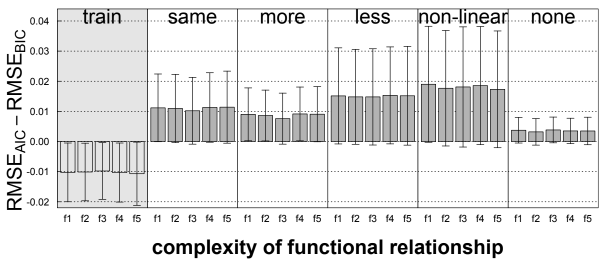
clust:PCA:sum



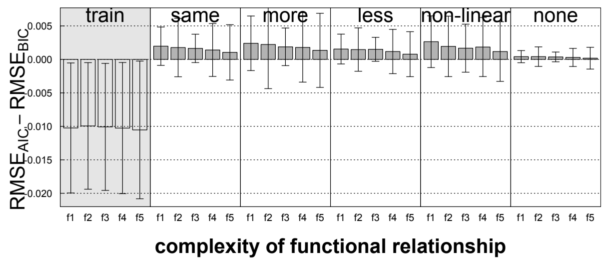
clust:PCA:ex



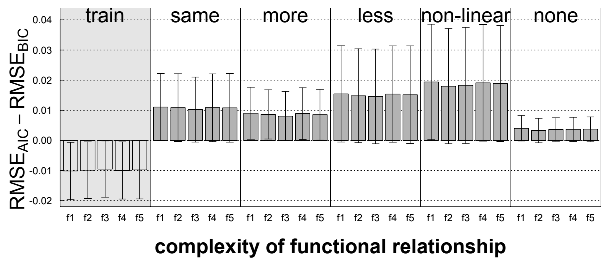
clust:clust:hoeffward:cent



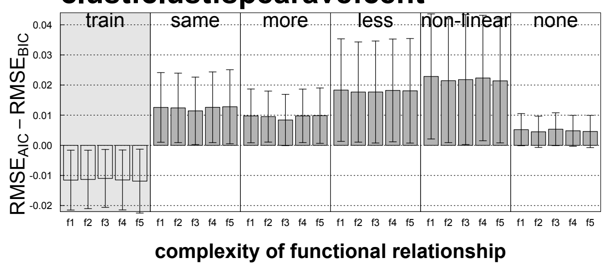
clust:clust:hoeffward:sum



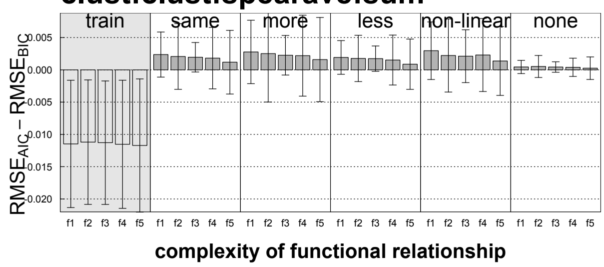
clust:clust:hoeffward:ex



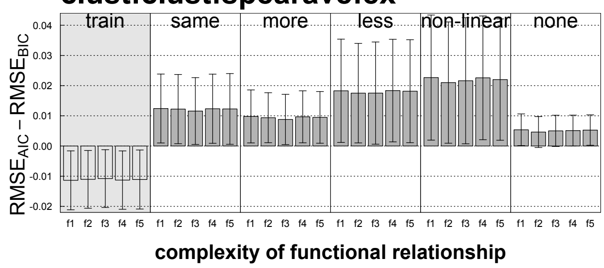
clust:clust:spearave:cent



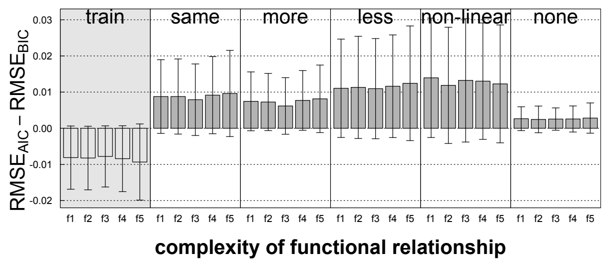
clust:clust:spearave:sum

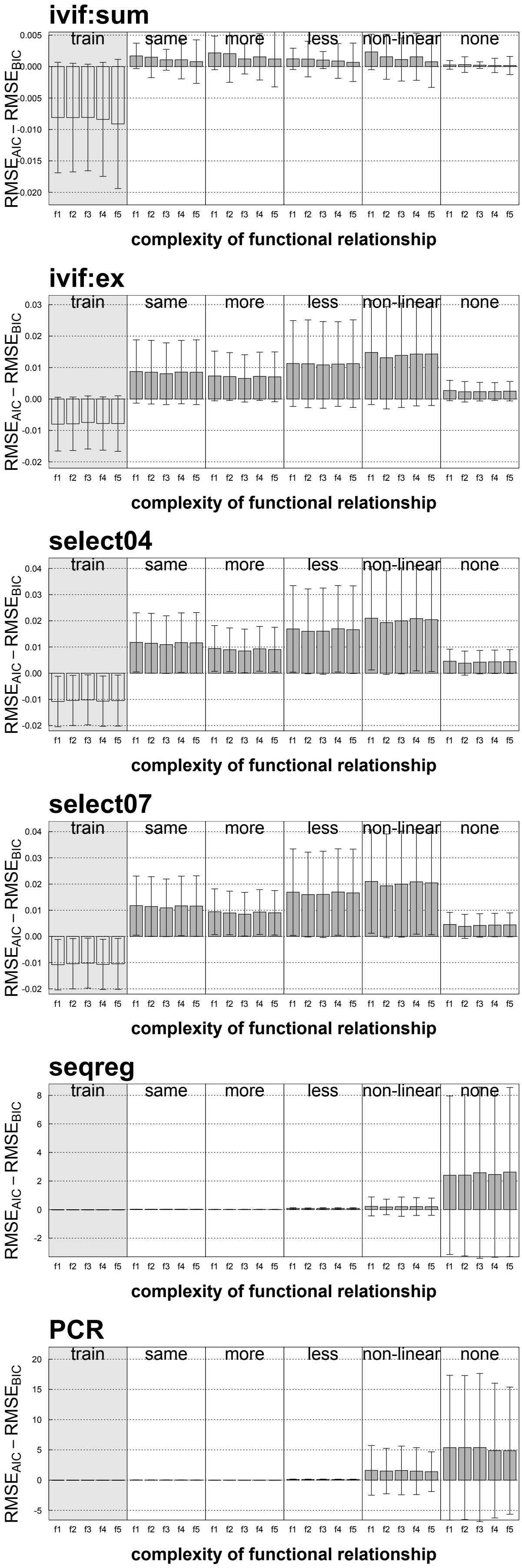


clust:clust:spearave:ex



ivif:cent





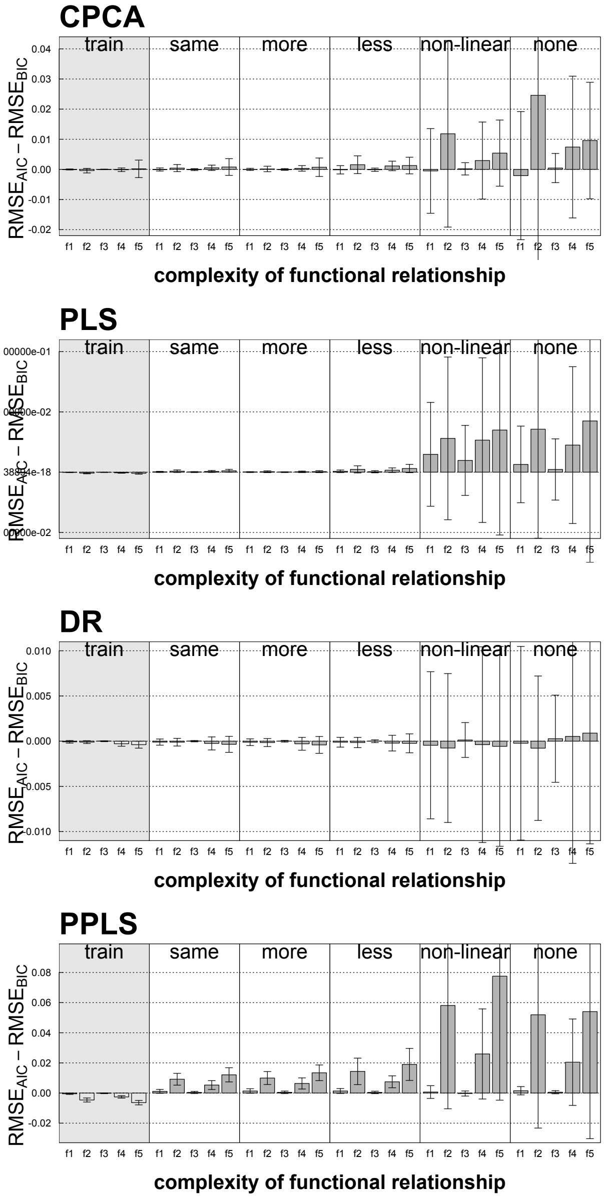


Fig. A2. Relative performance of AIC- and BIC-based model selection for all approaches (as measured by Root Mean Squared Error). Positive values indicate better performance of BIC (lower error on validation data). Note change in scale of y-axis. While overall BIC is superior, differences are small except for GLM, CWR, seqreg and PCR.

Pre-analysis: dealing with clusters

Each of the four ways to identify clusters (PCA, clustering based on Hoeffding-Ward, clustering based on Spearman-average and iVIF) was combined with three methods to extract a new representative variable for a cluster (central variable, first PC of cluster members, variable which best explains the response; Fig. A3). All clustering methods were rather similar in fit and prediction, although Spearman-average and iVIF tended to fare slightly better. While choosing the variable with the highest univariate correlation with the response ("expl") tended to be the best way to process clusters, it also is statistically least sound (sometimes referred to as "data snooping", because the response variable is examined during data reduction: Harrell 2001, p. 160). Using the central variable (which has the highest correlation with all other variables in the cluster) generally performed better than using the cluster's first PC (Fig. A3).

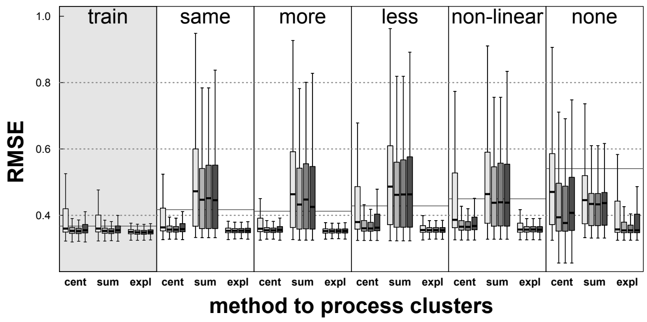


Fig. A3. Root mean square error of model predictions for training and test data sets. The four clustering methods (left to right in each group: PCA, clustering based on Hoeffding-Ward, clustering based on Spearman-average and iVIF), cluster processing methods are abbreviated as cent=central variable; sum=first PC of cluster members; and expl=variable explaining response best. Grey line is mean value for data set. Data points outside 1.5 x interquartile range omitted for clarity.

Method comparison across simulations – alternative measures of performance

In addition to the RMSE presented in the main text, we also analysed R^2, slope and intercept of the calibration curve. The results for these indices are given here, for test same and none only.

R^2

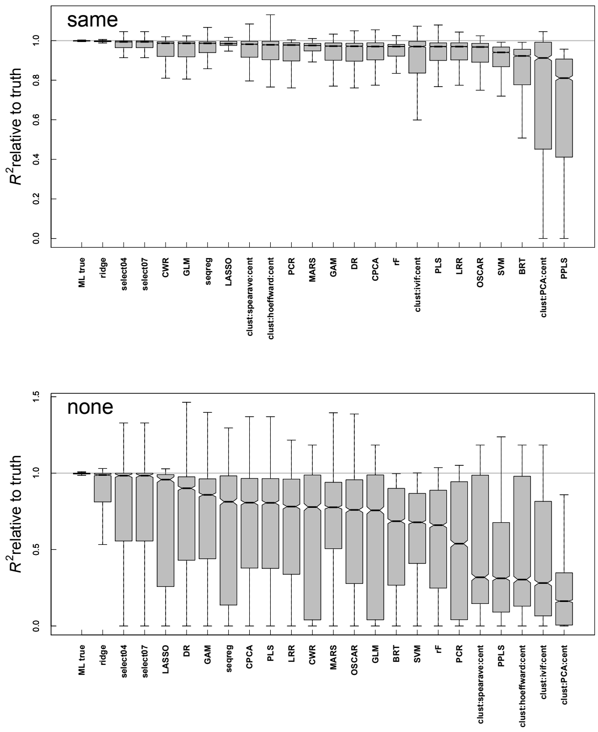


Fig. A4. Values are scaled such that the true model (ML true) has an R^2 of 1. Compare to sequence in Fig. 5 (main text).

Slope

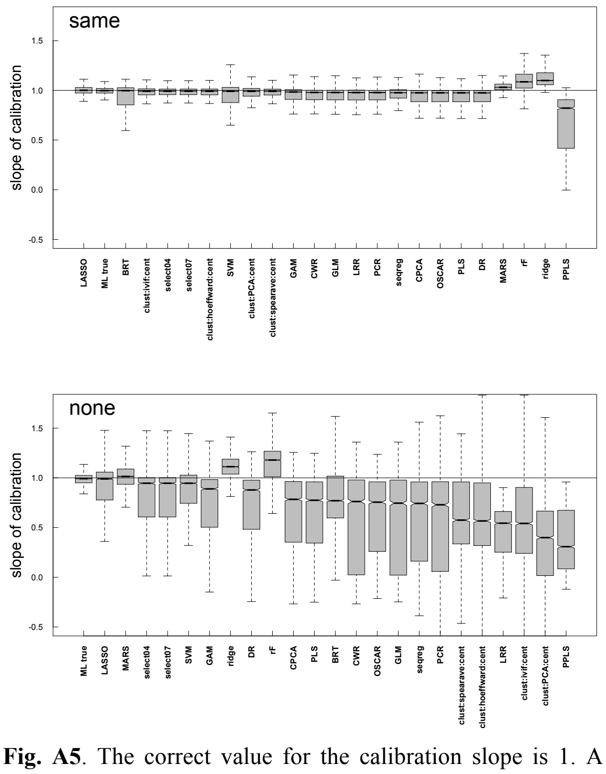


Fig. A5. The correct value for the calibration slope is 1. A value of 0 indicates a horizontal line, i.e. no correlation between observed and predicted values. Note that only ridge and randomForest yield slopes steeper than 1. Compare to sequence in Fig. 5 (main text).

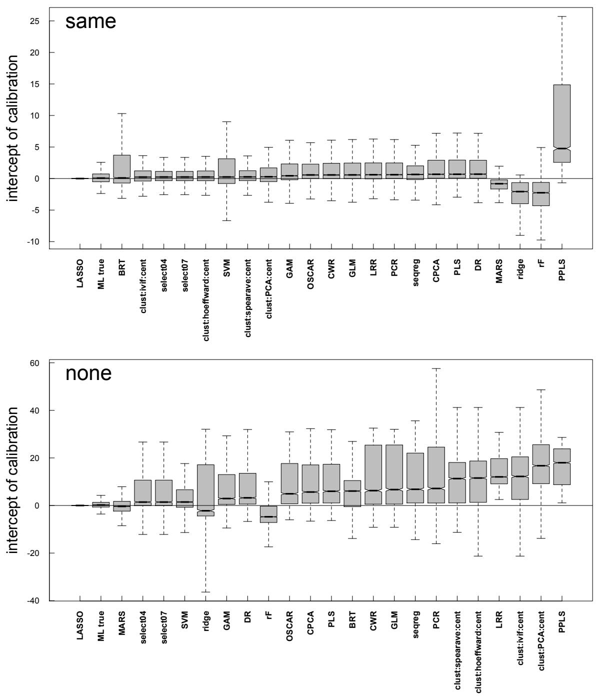


Fig. A6. The correct value for the calibration intercept is 0. Note that only ridge and randomForest yield intercepts less than 0, thereby correcting for the too steep slopes (Fig. A5). Intercepts larger than 0 indicate systematic under-prediction. Compare to sequence in Fig. 5 (main text).

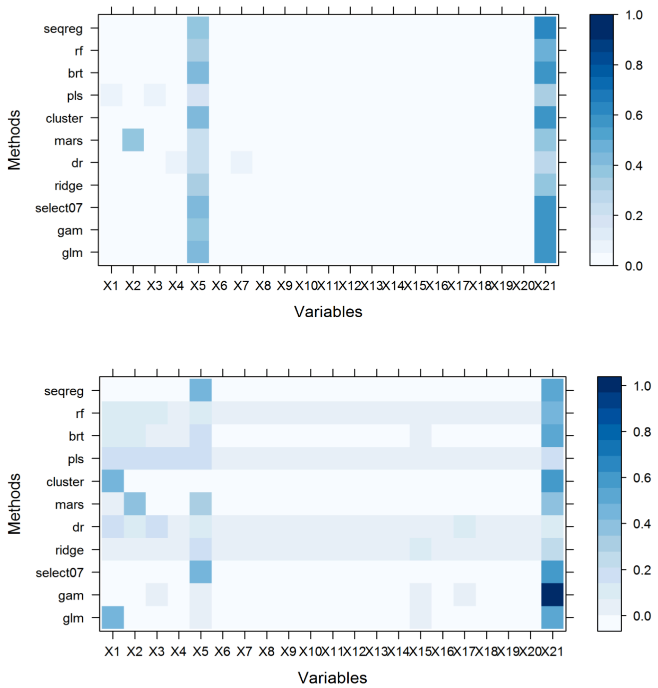


Fig. A7. Relative importance plots for simulated data. The upper part shows the relative importance of the variables for mild collinearity (decay = 8), while the lower part shows the relative importance for severe collinearity (decay = 3). The true functional relationship is: $y = 25 + X_5 + X_{21} + X_5^2 + X_{21}^2 + X_5 * X21$ (function 5). To compare the importance each method assigns to different explanatory variables we estimated the relative importance of each variable. Depending on the method, the Sum of Squares, deviance (seqreg), F-value (GAM), relative importance (BRT) or a factor calculated from the loadings and the coefficients (PLS) was used as an importance measure.

Intercept

Case studies

To see how the different methods behave under real world conditions, we applied a subset of the most promising methods to three case studies. These case studies serve to illustrate the different ecological conclusions that different approaches suggest.

Case study 1: Two simulated data sets

As an example for the effects of collinearity on model selection we used two data sets of the simulations. One data set shows severe collinearity while the other has been constructed under mild collinearity. For both data sets, model fit was high: all methods explained between 89 and 91 percent of the variance. But increasing collinearity led to problems for the identification of the right model structure for most methods (cf. Fig. A7). While most methods performed well under mild collinearity, only select07 and sequential regression were able to identify the correct model structure under severe collinearity. Notice that both methods use data snooping. GAMs did place too much importance on X_{21}, while both GLM and the Hoeffding/Ward cluster approach focused on X_1 instead of X_5. Machine learning approaches, PLS and DR incorrectly distributed the importance over many predictors. MARS worked well in general but erroneously took X_2 into the model in both cases.

Case study 2: Occurrence pattern of Black Grouse (*Tetrao tetrix*) in Europe

In this case study, we explored the relationship between European Black Grouse (*Tetrao tetrix*) distribution and observed climate (Fig. A8). Georeferenced species records were obtained from GBIF Data Portal (www.data.gbif.org, accessed 2011-02-02) and then aggregated to 2.5' resolution, resulting in a total of 12445 occurrence records. The same number of random absences was drawn from the background, restricted to Europe. Bioclimatic variables for current climate were complied from the WorldClim database at 2.5' resolution (Hijmans et al. 2005). For our analysis, we created random subsets of 5000 presence-absence records for model training and 5000 records for independent evaluation. Prior to analysis all predictor variables were standardised. Pairwise correlations between the 19 predictors were as high as |r|=0.95 (Fig. A9). Bio7 is defined through a linear combination of other variables and was hence omitted from subsequent analysis.

Model performance was evaluated against independent test data in terms of explanatory power (explained deviance R^2, Menard 2000) and in terms of discriminatory

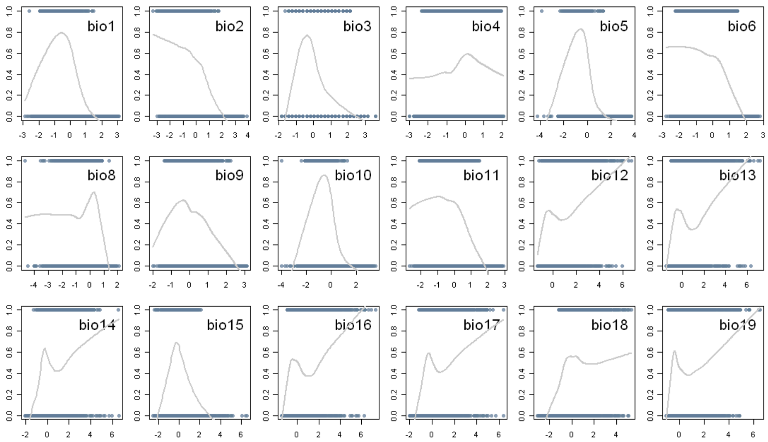


Fig. A8. Scatterplots of Black Grouse occurrence against each predictor. Lines are locally weighted smoothers.

power (AUC, area under the ROC curve, Fielding and Bell 1997). Most methods achieved excellent AUC scores (>0.9, Hosmer and Lemeshow 2000) and reasonable goodness of fit values R^2 between 0.50 and 0.70. RandomForest performed best according to both measures, followed by BRT and GAM (Fig. A10). Select07 achieved slightly lower accuracy values than the other methods (AUC=0.87, R^2=0.42), iVIF-based clustering performed worst (AUC=0.75, R^2=0.20).

By and large, variable importance was concordant between methods. Temperature related variables (bio1 – bio11) generally were deemed more important for Black Grouse occurrence than precipitation (bio12 - bio19; Fig. A10). Also, annual trends in temperature (bio1) and limiting temperatures (e.g. mean temperature of warmest quarter, bio10) were identified as more important than temperature seasonality (e.g. bio4). The clustering methods (Hoeffding-Ward and iVIF-based) and Select07 selected distinctly different variables than the rest of the methods as they aimed to find representative subsets of predictors. Thereby, iVIF-based clustering yielded ecologically implausible clusters while Hoeffding-Ward selected plausible clusters. Several regression-based methods (GLM, GAM, Ridge, PLS, MARS) estimated ecologically implausible relationships for some, generally less important predictors (e.g. bimodal response for bio6, minimum temperature of coldest month).

Overall, the methods compared provided a consistent picture of relative climate effects on Black Grouse occurrence in Europe. More flexible model types such as randomForests or BRTs performed best, although differences to classical methods were not very pronounced. However, when using species distribution models as predictive tool we need to be aware that high performance scores under current climate do not guarantee robustness under changing environmental conditions (Elith et al. 2010). Rather, this analysis underscores the need to scrutinize model plausibility.

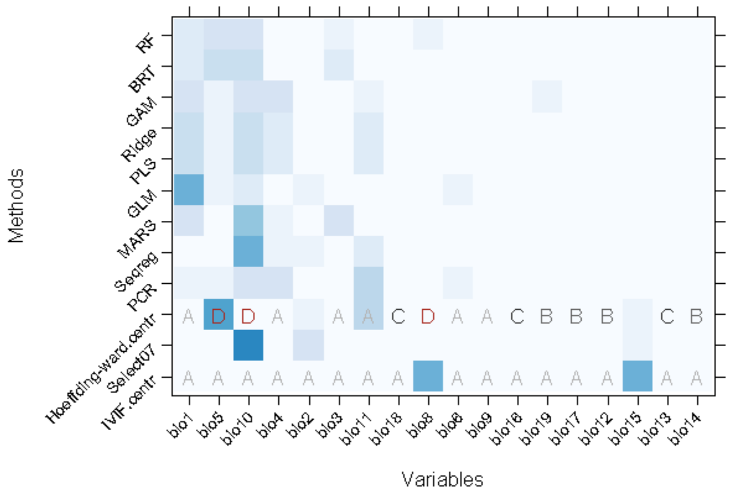


Fig. A10. Relative importance plot for Black Grouse case study. For clustering methods similar letters indicate that predictors belong to the same cluster. Methods are ranked according to the explained deviance. Predictor variables are ranked according to their mean rank assigned by all methods.

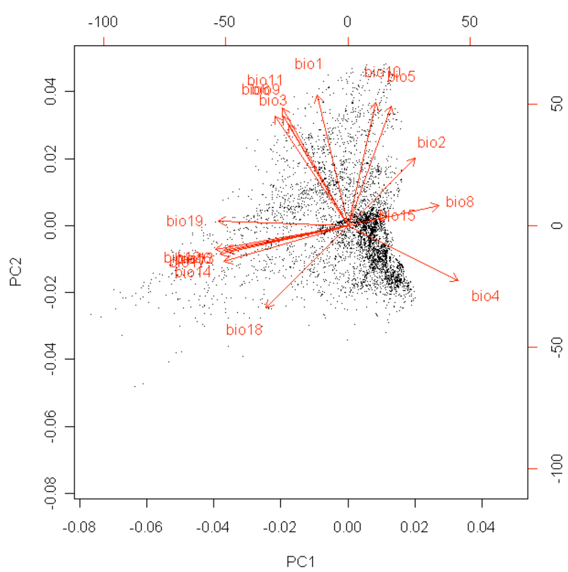


Fig. A9. PCA of predictors for Black Grouse.

Case study 3: Drivers of global bird diversity

This analysis used data on climate, terrain, land use and protected areas as well as economic and political indicators to analyse (normally distributed) global bird diversity (species density per km^2) at the country level for 224 countries (cf. Fig.s B11 and B12). Species-richness area-effects were corrected for by calculating the bird species density, i.e. residuals of a linear regression[1] of log-log transformed bird species richness and country area (yielding bird species/km^2). This correction ensured that the trivial effect of country size did not obscure the effects of other predictors. The species-

[1] The assumption of normality has been tested based on graphical procedures and statistical test.

area-relationship explained 60% of the variance in the

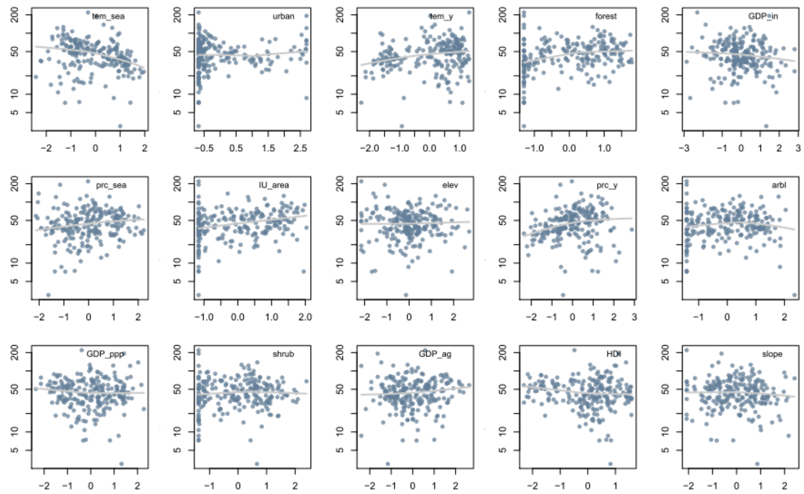


Fig. A11. Scatterplots for the corrected diversity of birds and the 15 most important predictors as selected by random forest. Lowess smoothers have been added to aid interpretation. The y-axis is log-transformed.

original data. Missing data points for explanatory variables were imputed based on all other predictors (Harrell 2001, p. 47). Explanatory variables were Box-Cox-transformed to reduce effects of skewed predictors and then standardised to combat trivial forms of collinearity (Quinn and Keough 2002, p. 131).

Variable names refer to temperature seasonality (temp_sea), proportion urban area, mean annual temperature (temp_y), percent forest, industrial gross domestic product (GDP_in), precipitation seasonality (prc_sea), area number of protected sites according to IUCN (IU_area), elevation (elev), annual precipitation (prc_sea and prc_y, respectively), percent arable land (arbl), GDP per person (GDP_ppp), percent shrubland (shrub), agricultural contribution to the GDP (GDP_ag), human development index according to the World Bank (HDI) and the average inclination of the country (slope).

Most methods achieved a reasonable goodness of fit with the best approaches yielding an R^2 of 0.59 (seqreg) and 0.50 (select07), going down to values below 0.30 (for ridge, cluster and randomForest).

There was no strong pattern in which variables were selected by the approaches. Urban area was the only variable included in all models, coming out second after temperature seasonality overall (Fig. A13). Several points are worth noting, however. Firstly, most methods kept many variables in the model, even though these had little to contribute (relative partial R^2-values of less than 0.1, noticeable as light blue in Fig. A13). As flip-side of this finding, only a few models clearly identified one or two variables as supremely important (values > 0.25: temperature seasonality by seqreg, select07, BRT and GLM; urban by cluster; voice/accountability by MARS; industrial GDP by GLM; governmental efficiency and political stability by DR). Secondly, it is ecologically implausible that GDP and the socio-political situation are causally related to bird species density in this data set. Still, these variables were selected repeatedly and by very different approaches. In the extreme, DR selected virtually only socio-political variables, while even the ecology-dominated seqreg model included voice/accountability. Thirdly, human-influenced plausible predictors were assigned weight very irregularly. The amount of urban area as well as the number and area of nature reserves shows up as relevant in some models, but not in others. Urban area can have both positive and negative effects on species density. Densely populated countries often have a higher density of species records, thereby increasing the apparent species richness in the data. At the same time, landscapes are usually more transformed there and hence habitats degraded, leading to lower population sizes and detection of bird species.

Overall, this analysis does not offer a consistent explanation for the observed global birds species density. Variables are too confounded to unambiguously evaluate the effect of temperature seasonality or nature reserves.

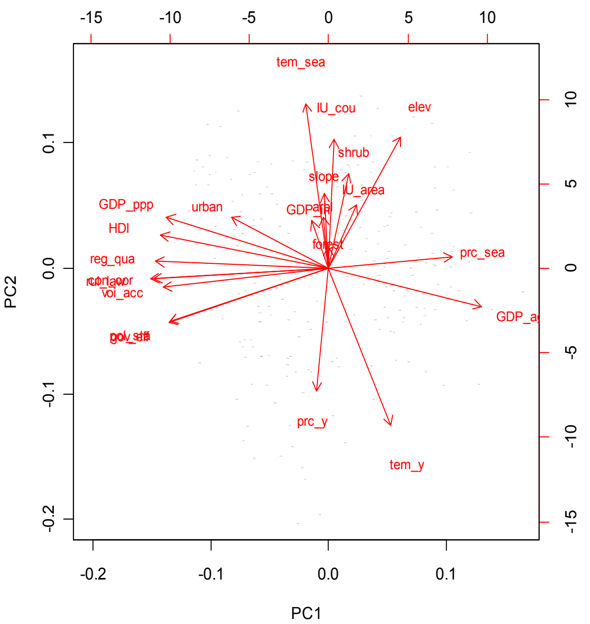


Fig. A12. PCA of predictors for bird diversity. See caption Fig.

Discussion

Our case studies were meant to illustrate the application of collinearity methods to real data sets. The problems we confront with real data, compared to our simulated data, are manifold: small sample size, extremely heterogeneous collinearities, categorical variables, non-normal response variables (case study 2), highly-skewed predictors and many more.

Case study 1, which looked at two of our 4000 simulated data sets, showed that some methods are substantially affected by high collinearity. In particular latent variable methods (PLS, DR), randomForest, BRT and ridge were unable to attribute effects correctly to predictors 5 and 21. These methods smeared the effect out over all (several: BRT) variables (although quanti-

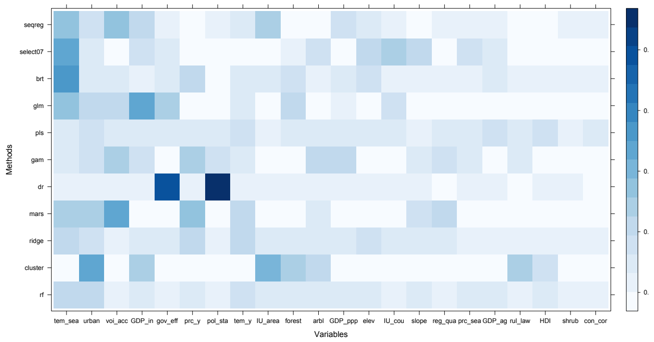


Fig. A13. Global bird diversity case study: relative importance of predictor in the different collinearity approaches. Variables sequence is according to their average relative importance across the 10 methods (decreasing from left to right); methods sequence according to their R^2 (decreasing from top to bottom). See Fig. A5 for names of variables.

fication of variable importance is no exact science and alternative approaches to do so may paint a slightly different picture).

In case study 2, Black Grouse distribution in Europe, we observed three types of variables: those selected by several approaches, those selected by only one approach, and finally those never selected (Fig. A10). Among the group of variables selected by several approaches were bio1, bio5, bio10, bio2, bio3 and bio11, all relating to temperature. Which of them was selected to represent climate effects differed between modelling approaches, however. Precipitation-related variables (bio12-19) were rarely selected, except by iVIF-clustering.

Our case study 3, the analysis of variables driving bird diversity in the countries of the world, highlighted the problem of non-concurrent model results. Findings here were very different to case study 2, since virtually every predictor was selected by one method or another. Most noticeable in this case study is that several methods gave many predictors some importance, thus smearing importance across all of them (seqreg, BRT, PLS, ridge, randomForest: Fig. A13). The BRT approach, which put importance on every variable, led to an excellent fit but came with the problem of interpretability. The other extreme was formed by the DR clustering approach, which resulted in a manageable number of predictors but at the cost of a rather poor fit: it selected two nonsensical socio-economic variables as key predictors (political stability and governmental efficiency). Using a large number of predictors did not necessarily lead to a good fit: ridge as well as randomForest distributed the importance over a large number of predictors but still yielded only moderate fits. Only sequential regression and select07 led to relatively small, easily interpretable models. Favouring parsimonious models in the presence of high collinearity might be misleading: interpretation needs to consider the collinear variables as well because the central variable of a proxy set still reflects the shared variance with the collinear predictors.

From our case studies we conclude that under severe collinearity all methods are likely to remain oblivious to the underlying true relationships. Under moderate collinearity, however, all methods performed acceptably. Our analyses cannot substantiate claims that some methods are inherently preferable to others, once the threshold to severe collinearity has been surpassed.

Appendix 1.3: Methods not incorporated in this study

Since the realm of regression methods is vast, we have focussed here on methods commonly used or likely to have promise; the review and following case study is not exhaustive. (All code for data generation is available in the supplementary material and interested readers can apply it to any method we failed to cover.) Most noticeably, we did not include **factor analysis**, although it is a commonly employed method in sociological and also ecological research (see Glass and Taylor 1966, Gorsuch 1993, for classical accounts). Factor analysis is an exploratory or confirmatory technique, not primarily designed to remove collinearity in predictors. If used in that way, results largely fall together with those of principal component analysis. Furthermore, in factor analysis the number of factors to be extracted needs to be optimised. During preparatory analyses, factor analyses failed to converge both under very high and very low collinearity, and hence proved difficult to automate (Velicer and Douglas 1990, Velicer and Jackson 1990).

We also omitted **hierarchical partitioning** (Chevan and Sutherland 1991, Heikkinen et al. 2005), because although it helps to separate individual and joint effects of correlated variables, it lacks statistical rigor (there is, e.g., no statistical basis for the averaging across hierarchies). Also, the models specified are over-simplified since hierarchical partitioning models do not allow for non-linear relationships between response and predictor, nor interactions between predictor variables.

One major theme in machine-learning is to combine multiple models into one (Hastie et al. 2009). In analogy, multi-model predictions based, e.g., on AIC-weights of alternative (generalised) linear models (Burnham and Anderson 2002) could also form a basis for a method unaffected by collinearity in the sense that different predictors could be assigned to different models. While this is an option for small data sets with few predictors (see e.g. Dormann et al. 2008b, for an application), large data sets and many predictors require huge computer resources with little gain over efficient machine-learning approaches, and hence we did not investigate this option further.

Two other methods we knowingly omitted were **path analysis**, because it requires an expert pre-definition of possible models and is hence not amenable to automatic implementation; and the **stepwise collinearity diagnostic** (SCD: Brauner and Shacham 1998), because we found no open source implementation.

Literature cited

Bondell, H. D. and B. J. Reich. 2007. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. Biometrics **64**:115-121.

Booth, G. D., M. J. Niccolucci, and E. G. Schuster. 1994. Identifying proxy sets in multiple linear regression: an aid to better coefficient interpretation. Res. Pap. INT-470, US Dept. of Agriculture, Forest Service, Ogden, UT.

Bortz, J. 1993. Statistik für Sozialwissenschaftler. Springer Berlin.

Brauner, N. and M. Shacham. 1998. Role of range and precision of the independent variable in regression of data. American Institute of Chemical Engineers Journal **44**:603-611.

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

Burgess, C. 1998. A tutorial on support vector machines for pattern recognition. Knowledge Discovery and Data Mining **2**:121-167.

Burnham, K. P. and D. R. Anderson. 2002. Model Selection and Multi-Model Inference: A Practical Information-Theoretical Approach. Springer, Berlin.

Capen, D. E., J. W. Fenwick, D. B. Inkley, and A. C. Boynton. 1986. Multivariate models of songbird habitat in New England forests. Pages 171-175 *in* J. A. Verner, M. L. Morrison, and C. J. Ralph, editors. Wildlife 2000: Modelling Habitat Relationships of Terrestrial Vertebrates. University of Wisconsin Press, Madison, WI.

Carrascal, L. M., I. Galván, and O. Gordo. 2009. Partial least squares regression as an alternative to current regression methods used in ecology. Oikos **118**:681-690.

Chevan, A. and M. Sutherland. 1991. Hierarchical partitioning. American Statistician **45**:90-96.

Cook, R. D. and S. Weisberg. 1999. Applied Regression Including Computing and Graphics. Wiley, New York.

Cutler, D. R., J. T. C. Edwards, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler. 2007. Random Forest for classification in ecology. Ecology **88**:2783-2792.

Dormann, C. F., O. Purschke, J. R. García Márquez, S. Lautenbach, and B. Schröder. 2008a. Components of uncertainty in species distribution analysis: A case study of the Great Grey Shrike *Lanius excubitor* L. Ecology **89**:3371-3386.

Dormann, C. F., O. Schweiger, P. Arens, I. Augenstein, S. Aviron, D. Bailey, J. Baudry, R. Billeter, R. Bugter, R. Bukácek, F. Burel, M. Cerny, R. D. Cock, G. D. Blust, R. DeFilippi, T. Diekötter, J. Dirksen, W. Durka, P. J. Edwards, M. Frenzel, R. Hamersky, F. Hendrickx, F. Herzog, S. Klotz, B. Koolstra, A. Lausch, D. L. Coeur, J. Liira, J. P. Maelfait, P. Opdam, M. Roubalova, A. Schermann-Legionnet, N. Schermann, T. Schmidt, M. J. M. Smulders, M. Speelmans, P. Simova, J. Verboom, W. K. R. E. van Wingerden, and M. Zobel. 2008b. Prediction uncertainty of environmental change effects on temperate European biodiversity. Ecology Letters **11**:235-244.

Elith, J. and J. Leathwick. 2007. Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. Diversity and Distributions **13**:265-275.

Elith, J., J. R. Leathwick, and T. Hastie. 2008. A working guide to boosted regression trees. Journal of Animal Ecology **77**:802-813.

Evritt, B. S., S. Landau, and L. Morven. 2001. Cluster Analysis 4th edition. Hodder Arnold, London.

Fielding, A. H. and P. F. Haworth. 1995. Testing the generality of bird-habitat models. Conservation Biology **9**:1466-1481.

Friedman, J. F. and J. J. Meulman. 2003. Prediction with multiple additive regression trees with application in epidemiology. Statistics in Medicine **22**:1365-1381.

Friedman, J. H. 1991. Multivariate adaptive regression splines. Annual Statistics **19**:1-141.

Friedman, J. H. 2002. Stochastic gradient boosting. Computational Statistics & Data Analysis **38**:367-378.

Friedman, J. H., T. Hastie, and R. Tibshirani. 2000. Additive logistic regression: a statistical view of boosting. Annals of Statistics **28**:337-407.

Glass, G. and P. A. Taylor. 1966. Factor analytic methodology. Review of Educational Research **36**:566-587.

Gorsuch, R. L. 1993. Factor Analysis. 2nd edition. Lawrence Erlbaum, Hillsdale, NJ.

Grace, J. B. 2006. Structural Equation Modeling and Natural Systems. Cambridge University Press, Cambridge.

Graham, M. H. 2003. Confronting multicollinearity in ecological multiple regression. Ecology **84**:2809-2815.

Green, R. H. 1979. Sampling Design and Statistical Methods for Environmental Biologists. Wiley, New York.

Guo, Q. H., M. Kelly, and C. H. Graham. 2005. Support vector machines for predicting distribution of sudden oak death in California. Ecological Modelling **182**:75-90.

Hastie, T., R. Tibshirani, and J. H. Friedman. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd edition. Springer, Berlin.

Heikkinen, R. K., M. Luoto, M. Kuussaari, and J. Pöyry. 2005. New insights into butterfly–environment relationships using partitioning methods. Proceedings of the Royal Society **272**:2203-2210.

Heinänen, S. and M. v. Numers. 2009. Modelling species distribution in complex environments: an evaluation of predictive ability and reliability in five shorebird species. Diversity and Distributions **15**:266-279.

Hernandez, P. A., C. H. Graham, L. L. Master, and D. L. Albert. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography **29**:773-785.

Hoerl, A. E. and R. W. Kennard. 1970. Ridge regression: biased estimation for non-orthogonal problems. Technometrics **12**:55-67.

Jackson, J. E. 1991. A User's Guide to Principal Components. Wiley, New York.

Kaufman, L. and P. J. Rousseeuw. 2005. Finding Groups in Data: An Introduction to Cluster Analysis. 2nd edition. Wiley, New York.

Krämer, N., A. L. Boulesteix, and G. Tutz. 2007. Penalized partial least squares with applications to B-splines transformations and functional data. preprint available at http://ml.cs.tu-berlin.de/~nkraemer/publications.html.

Leathwick, J. R., J. Elith, M. P. Francis, T. Hastie, and P. Taylor. 2006a. Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. Marine Ecology Progress Series **321**:267-281.

Leathwick, J. R., J. Elith, and T. Hastie. 2006b. Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. Ecological Modelling **199**:188.

Martens, H. and T. Naes. 1989. Multivariate Calibration. Wiley, New York.

Moisen, G. G. and T. S. Frescino. 2002. Comparing five modelling techniques for predicting forest characteristics. Ecological Modelling **157**:209-225.

Oppel, S., H. M. Schaefer, V. Schmidt, and B. Schröder. 2004. Habitat selection by the Pale-headed brush-finch, *Atlapetes pallidiceps*, in southern Ecuador: implications for conservation. Biological Conservation **118**:33-40.

Pearce-Higgins, J. W. and D. W. Yalden. 2004. Habitat selection, diet, arthropod availability and growth of a moorland wader: the ecology of European Golden Plover *Pluvialis apricaria* chicks. Ibis **146**:335-346.

Pillar, V. D. 1999. How sharp are classifications? Ecology **80**:2508-2516.

Rawlings, J. O. 1988. Applied Regression Analysis: A Research Tool. Wadsworth & Brooks, Pacific Grove, California.

Reineking, B. and B. Schröder. 2006. Constrain to perform: regularization of habitat models. Ecological Modelling **193**:675-690.

Ribeiro, R. and L. Torgo. 2008. A comparative study on predicting algae blooms in Douro River, Portugal. Ecological Modelling **212**:86-91.

Riffell, S. K. and K. J. Gutzwiller. 2009. Interannual variation in birdlandscape relations: understanding sources of a pervasive conservation dilemma. Oikos **118**:45.

Schapire, R. E. 2002. The boosting approach to machine learning: an overview. MSRI Workshop on Nonlinear Estimation and Classification **in press**.

Schapire, R. E. 2003. The boosting approach to machine learning: An overview.*in* D. D. Denison, M. H. Hansen, C. Holmes, B. Mallick, and B. Yu, editors. Nonlinear Estimation and Classification. Springer, Berlin.

Schölkopf, B. and A. J. Smola. 2000. New support vector algorithms. Neural Computation **12**:1207-1245.

Steinwart, I. and A. Chistmann. 2008. Support Vector Machines. Springer, Berlin.

Tabachnick, B. and L. Fidell. 1989. Using Multivariate Statistics. Harper & Row Publishers, New York.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society B **58**:267-288.

Vapnik, V. 1996. The Nature of Statistical Leaning Theory. Wiley, New York.

Velicer, W. F. and C. Douglas. 1990. Component Analysis versus Common Factor Analysis II: some further observations. Multivariate Behavioral Research **25**:97-114.

Velicer, W. F. and D. A. Jackson. 1990. Component Analysis versus Common Factor Analysis I: some issues in selecting an appropriate procedure. Multivariate Behavioral Research **25**:1-28.

Vigneau, E. and E. M. Qannari. 2002. A new algorithm for latent root regression analysis. Computational Statistics & Data Analysis **41**:231-242.

Vigneau, E., E. M. Qannari, and D. Bertrand. 2002. A new method of regression on latent variables. Application to spectral data. Chemometrics and Intelligent Laboratory Systems **63**:7-14.

Weisberg, S. 2008. dr: Methods for dimension reduction for regression. R package version 3.0.3, URL http://www.r-project.org.

Yen, P. P. W., W. J. Sydeman, and K. D. Hyrenbach. 2004. Marine bird and cetacean associations with bathymetric habitats and shallow-water topographies: implications for trophic transfer and conservation. Journal of Marine Systems **50**:79-99.

Zou, H. and T. Hastie. 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67**:301-320.