Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2015

doi: 10.1111/2041-210X.12378

Mapping beta diversity from space: Sparse Generalised Dissimilarity Modelling (SGDM) for analysing high-dimensional data

Pedro J. Leitão[1,2]*, Marcel Schwieder[1], Stefan Suess[1], Inês Catry[2], Edward J. Milton[3], Francisco Moreira[2], Patrick E. Osborne[4], Manuel J. Pinto[5], Sebastian van der Linden[1] and Patrick Hostert[1]

[1]*Geography Department, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany;* [2]*CEABN – Centre for Applied Ecology "Prof. Baeta Neves"/InBio – Research Network in Biodiversity and Evolutionary Biology, Institute of Agronomy, University of Lisbon, Tapada da Ajuda, 1349 – 017 Lisbon, Portugal;* [3]*Geography and Environment, University of Southampton, Highfield, Southampton, SO17 1BJ, UK;* [4]*Centre for Environmental Sciences, Faculty of Engineering and the Environment, University of Southampton, Highfield, Southampton, SO17 1BJ, UK; and* [5]*Botanic Garden, National Museum of Natural History and Science (MNHNC), University of Lisbon, Rua da Escola Politécnica, 58, 1250-102 Lisbon, Portugal*

Summary

1. Spatial patterns of community composition turnover (beta diversity) may be mapped through generalised dissimilarity modelling (GDM). While remote sensing data are adequate to describe these patterns, the often high-dimensional nature of these data poses some analytical challenges, potentially resulting in loss of generality. This may hinder the use of such data for mapping and monitoring beta-diversity patterns.

2. This study presents Sparse Generalised Dissimilarity Modelling (SGDM), a methodological framework designed to improve the use of high-dimensional data to predict community turnover with GDM. SGDM consists of a two-stage approach, by first transforming the environmental data with a sparse canonical correlation analysis (SCCA), aimed at dealing with high-dimensional data sets, and secondly fitting the transformed data with GDM. The SCCA penalisation parameters are chosen according to a grid search procedure in order to optimise the predictive performance of a GDM fit on the resulting components. The proposed method was illustrated on a case study with a clear environmental gradient of shrub encroachment following cropland abandonment, and subsequent turnover in the bird communities. Bird community data, collected on 115 plots located along the described gradient, were used to fit composition dissimilarity as a function of several remote sensing data sets, including a time series of Landsat data as well as simulated EnMAP hyperspectral data.

3. The proposed approach always outperformed GDM models when fit on high-dimensional data sets. Its usage on low-dimensional data was not consistently advantageous. Models using high-dimensional data, on the other hand, always outperformed those using low-dimensional data, such as single-date multispectral imagery.

4. This approach improved the direct use of high-dimensional remote sensing data, such as time-series or hyperspectral imagery, for community dissimilarity modelling, resulting in better performing models. The good performance of models using high-dimensional data sets further highlights the relevance of dense time series and data coming from new and forthcoming satellite sensors for ecological applications such as mapping species beta diversity.

Key-words: biodiversity, community modelling, EnMAP, generalised dissimilarity modelling, hyperspectral data, Landsat, remote sensing, sparse canonical correlation analysis, time-series, turnover

Introduction

Recent global reduction in biodiversity is widely acknowledged, with direct impacts on ecosystem functioning and its provisioning of services (Cardinale *et al.* 2012). However, existing patterns of biodiversity and most particularly those of community composition turnover, or beta diversity, are little known (Ferrier *et al.* 2002; McKnight *et al.* 2007). A deeper knowledge of these patterns can provide insights into the ecological processes determining species and community distributions, such as the identification of ecological tipping points or of vulnerable taxonomic groups (Guerin, Biffin & Lowe 2013). This can also support well-informed management practices for mitigating biodiversity declines. While beta diversity is not a new concept (Whittaker 1960) and closely relates to that of ecological complementarity (Faith *et al.* 2003), its importance has received growing attention, particularly due to its implications for biodiversity conservation and ecosystem functioning (Hooper *et al.* 2005; Legendre, Borcard & Peres-Neto 2005).

*Correspondence author. E-mail: p.leitao@geo.hu-berlin.de

Many studies have dealt with the description of beta diversity and its measurement. A commonly used approach is one of data ordination, such as canonical correlation analysis (Legendre, Borcard & Peres-Neto 2005). In this approach, the community data are transformed by incorporating environmental variables of interest as constraints for the ordination, which also allows the inference of species–environment relationships (Legendre & Gallagher 2001). Another common approach for analysis of beta diversity is through dissimilarity measures of the community data (Ferrier *et al.* 2007; Tuomisto 2010; De Caceres, Legendre & He 2013). Ferrier *et al.* (2007) introduced an approach called generalised dissimilarity modelling (GDM), which is suitable for modelling and mapping spatial patterns of community composition turnover. In this approach, the compositional dissimilarity between all pairs of samples is modelled as a function of environmental distance, using a linear combination of I-spline basis functions. The model architecture constrains the fitted functions to be monotonic, with the assumption that increasing separation of sites along an environmental gradient can only result in increasing compositional dissimilarity (Ferrier *et al.* 2007). The spatial pattern in community compositional change predicted by GDM can then be visualised through the nonlinear ordination of the predicted dissimilarities between location pairs.

Remotely sensed data, by repeatedly describing the Earth's surface in a synoptic and detailed manner, are suitable for monitoring ecological processes (Kerr & Ostrovsky 2003; Turner *et al.* 2003). The global extent and timely coverage of these data make them particularly suitable for continuous large area ecosystem monitoring (Griffiths *et al.* 2012; Hansen *et al.* 2013). Moreover, the opening of the Landsat data archive and the advent of new global monitoring satellites, such as NASA's Landsat 8 (operational since May 2013), the European Space Agency's Sentinel missions (launches due between 2013 and 2015) and the German hyperspectral EnMAP mission (launch due in 2017), further enhances the potential of this data source (Kennedy *et al.* 2014). While choosing the right remote sensing data or product is not always an easy matter (Cord *et al.* 2013), making full use of the continuous information of such data (i.e. unclassified remote sensing data or derived products) has been shown to be advantageous in several studies on species distributions (Osborne, Alonso & Bryant 2001; Parviainen *et al.* 2013; Cord *et al.* 2014). Indeed the spatial variation of the reflection signal closely describes the spatial patterns of vegetation and other landscape features which might determine species occurrence and abundance patterns. Measures of heterogeneity and distance of remotely sensed spectra have been successfully used for characterising species alpha and beta diversities (Rocchini 2007; Feilhauer & Schmidtlein 2009; Rocchini *et al.* 2010; Baldeck & Asner 2013). On the other hand, the high-dimensional (and potentially multicollinear) nature of these data poses challenges for their analysis (Dormann *et al.* 2013), potentially resulting in lack of performance and generality.

An advance in dealing with high-dimensional data sets is sparse canonical correlation analysis (SCCA; Witten, Tibshirani & Hastie 2009), a form of regularised ordination. This method stems from genetics research where the number of variables is typically much greater than the number of samples (Witten & Tibshirani 2009), which parallels the analysis of high-dimensional remote sensing data. SCCA is based on the least absolute shrinkage and selection operator or LASSO (Tibshirani 1996), a regularisation approach aimed at optimising performance while reducing model complexity through penalisation (Reineking & Schröder 2006). In the LASSO regression, the sum of the absolute values (L1-norm) of the parameter estimates is used for penalisation, which encourages sparse solutions via shrinkage of coefficients towards zero, effectively selecting features (Tibshirani 1996; Tibshirani *et al.* 2005).

In this study, we present a methodological approach for improving the usage of GDM for fitting patterns of beta diversity, by addressing the issues of high-dimensionality data when using (unclassified) spaceborne spectral data. This method consists of fitting sparse canonical components (extracted through a SCCA) in a GDM, hereafter referred to as Sparse Generalised Dissimilarity Modelling or SGDM.

We tested this approach using data from a Mediterranean region in southern Portugal, where a spatial and environmental gradient of shrub encroachment following land abandonment results in a progressive transition from open farmland fields to dense shrublands and forests (Moreira *et al.* 2007). This encroachment affects the structure and functioning of the ecosystem (Eldridge *et al.* 2011), including the compositional turnover in the existing bird communities (Leitão, Moreira & Osborne 2010).

The predictive performance of SGDM was compared with that of GDM using several high- and low-dimensional remote sensing data sets, including single date and time series of multispectral Landsat TM data and (simulated) hyperspectral EnMAP data. All code necessary to run the presented approach is provided (see Data S1), including several general GDM tools (e.g. the calculation of variable contribution significance, and the leave-one-out cross-validated performance), and some specific SGDM functions.

Materials and methods

SPARSE GENERALISED DISSIMILARITY MODELLING

The SGDM approach requires the input of two data matrices, one of species occurrence or abundance data and one of environmental variables, in a canonical correspondence analysis manner. It consists of initially transforming (and in this way reducing) high-dimensional environmental data by means of a SCCA (Witten, Tibshirani & Hastie 2009; Fig. 1), in order to maximise the correlation between transformed environmental and species data. The SCCA, being a form of penalised canonical correlation analysis, applies the L_1 (lasso) penalty function on the data matrices to resolve the sparse canonical vectors which can then be applied to ordinate the data. The penalty to be applied to each data matrix (the L_1 bound on the respective canonical vector) is in the form

```
c_1||u||_1ncol(x)forx,
c_2||v||_1ncol(y)fory,
```


which assumes values between 0 and 1 (larger L_1 bound corresponds to less penalisation) and *ncol* is the number of columns of the input matrix *x*. The SCCA requires the definition of two penalisation parameters, one for each of the data matrices (species and environmental). In SGDM, these are chosen via a heuristic grid search of all possible penalisation parameter pair combinations, in order to maximise the resulting GDM predictive performance. Effectively, for each penalisation pair combination, the resulting sparse canonical components are extracted and subsequently used for GDM, and the respective model performance inspected in a leave-one-out cross-validation procedure (i.e. by leaving out one site and all corresponding site pairs at each time). The parameter pair which results in higher GDM performance (in the form of the lowest root-mean-square error) is then selected, and the resulting components used for further GDM analysis. All analyses were run in R (R Development Core Team 2013) using several packages as described below.

In the proposed implementation of the SCCA parameterisation, which is run with the package PMA (Witten, Tibshirani & Hastie 2009), the type of data is set as 'standard' (for unordered data columns), a default 0·1 incremental step is given for the parameter grid search (although this can be manually defined), and the analysis is repeated in 50 iterations for algorithmic convergence. The number of sparse components to be extracted needs to be defined a priori, which we set as the maximum number of possible components, that is the minimum number of columns (species or environmental variables) between both matrices. The GDM model is run with the packages GDM4TABLES (freely available at https://sites.google.com/site/gdmsoftware/) and additional code from the package GDM01, under development at the R-Forge SCM repository (Ferrier *et al.* 2007). The dissimilarity metric to be used in the GDM needs to be defined. Here we used the default Bray–Curtis dissimilarity (Bray & Curtis 1957), which is widely used for count data.

The following step in the proposed approach is one of data reduction, to assure model parsimony. This is done by testing the significance of the input variable (sparse components) contribution, through matrix permutation, subsequently eliminating the non-significant variables (Ferrier *et al.* 2007). This step makes use of the packages GDM4TABLES, GDM01, VEGAN (Oksanen *et al.* 2012) and ECODIST (Goslee & Urban 2007).

For the purpose of beta-diversity mapping, the final GDM model can be applied to predict the dissimilarities between all sample pairs, and the predicted dissimilarities transformed to summarise most of the variability into few dimensions. The resulting transformed data can then be plotted in a map representing the patterns of community turnover (Ferrier *et al.* 2007).

CASE STUDY

In order to demonstrate the SGDM approach, we tested it on a study site around the towns of Castro Verde and Mértola in southern Portugal, along a gradient of shrub encroachment and subsequent bird community transition (Fig. 2). Extensive traditional agricultural practices in the region result in typical pseudo-steppe landscapes. These are characterised by dominant fallow grasslands, usually grazed by sheep (Moreira 1999), and a spatio-temporal mosaic of winter cereal crops, ploughed and stubble fields. Scattered rockrose (*Cistus* sp.) shrub patches are also common, mostly associated with rock outcrops or areas covered by shallow or skeletal soils and with the river valleys, as well as some areas of sparse, savanna-like holm oak (*Quercus rotundifolia*) woodlands. Agricultural land abandonment, however, has led to increasing shrub encroachment on fallow lands, which is particularly notable in the south-east of the study area (Schwieder *et al.* 2014). In contrast, the north-western half of the area lies within a designated Special Protection Area (SPA) for birds, where a directed agri-environmental scheme sets land-use incentives to keep traditional agricultural practices. This fosters the conservation of the local biodiversity, in particular a steppe bird community (Moreira *et al.* 2007), thus helping to maintain the pseudo-steppe mosaic within the SPA.

By having strong habitat associations, the existing bird communities are directly affected by changes in the landscape (Leitão, Moreira & Osborne 2010; Moreira *et al.* 2012). The observed gradient of increasing shrub encroachment, while potentially having beneficial effects on several ecosystem functions (e.g. soil protection against desertification; Marta-Pedroso *et al.* 2007; Eldridge *et al.* 2011), also results in a turnover of the bird assemblage composition, from the steppe bird community to one typical of Mediterranean shrublands (Moreira & Russo 2007; Leitão, Moreira & Osborne 2010).

We thus propose to model and map the region's bird community turnover along the shrub encroachment gradient by using a purposively collected species matrix and several high- and low-dimensional (remote sensing) environmental data sets, as described below.

DATA

Bird community data were collected in April 2011, according to a stratified sampling scheme, capturing a good geographical and successional representation of the study area (Leitão, Moreira & Osborne 2011). For this purpose, we defined six different landscape structural classes, with varying degrees of composition and configuration of woody vegetation, this way characterising the existing shrub encroachment gradient, from grasslands to fully established shrublands with successional tree cover. We also split the study region into geographical sections to ensure that all structural classes were covered on all sections, thus gua-

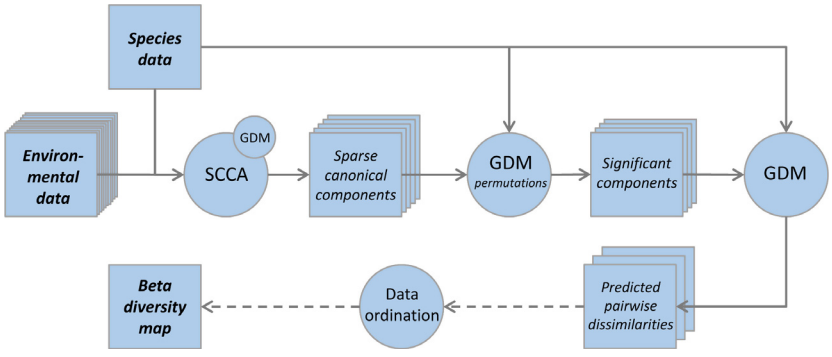

Fig. 1. Schematic workflow of the presented approach for the reduction of remote sensing data through a sparse canonical correlation analysis for generalised dissimilarity modelling. The resulting predicted dissimilarities can be subject to a data ordination for generating a beta-diversity map (shown with the dashed lines).

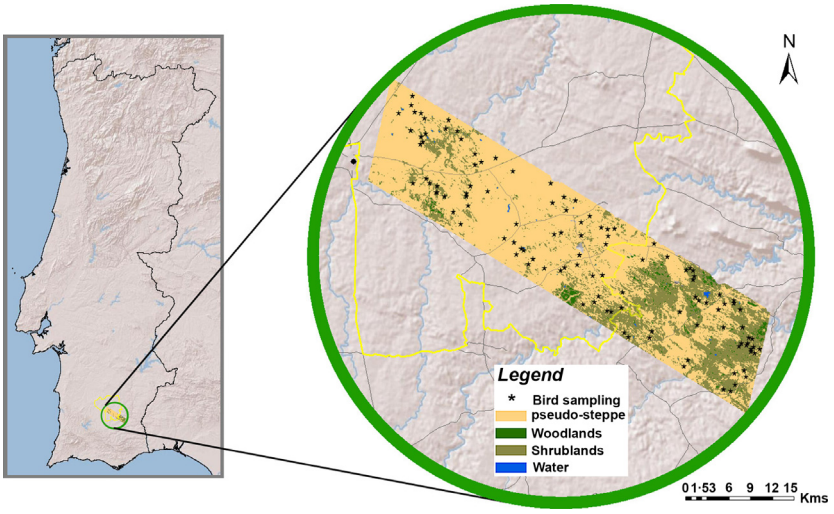

Fig. 2. Study area, including the bird sampling locations (black stars), the Castro Verde Special Protection Area limits (yellow line) and the land cover. The town of Castro Verde is marked with a black circle.

ranteeing a good representativeness of the variability found (Fig. 2). Bird assemblages were sampled using 10-min duration counts on circular plots with a 125 m distance limit (Fuller & Langslow 1984). All bird censuses were carried out during the birds' period of peak-activity, that is the early morning (first 4 h after sunrise) and evening (last 2 h before sunset) during the breeding season, and all visual and auditory bird observations were registered. Bird species not directly using the relevant (grassland to shrubby) habitats or those for which the sampling was not adequate (e.g. most raptors or aquatic birds) were excluded from the analysis. In total, 42 species were considered for modelling (see Table S1).

Several remote sensing data sets were used as environmental data to be tested with GDM and SGDM. We used a time series of Landsat-5 Thematic Mapper (TM) data from the year of 2011, acquired on six different dates between January and September (Julian dates 31, 79, 143, 175, 207 and 255) over our study area (path/row: 203/34; United States Geological Survey 2013). Only the six optical bands of the TM sensor were considered. All data were standard terrain corrected (L1T), and were further subject to radiometric and atmospheric correction using the Landsat Ecosystem Disturbance Adaptive Processing System (LEDAPS) algorithm (Masek *et al.* 2006). Both the time series (high dimensional) and the individual single-date (low dimensional) data were used for modelling. We also used simulated EnMAP (high dimensional) hyperspectral data (Stuffler *et al.* 2007; Segl *et al.* 2012), based on highly resolved airborne hyperspectral data (400–2500 nm) acquired in April and August of 2011 (Julian dates 097 and 223) over the study region (Schwieder *et al.* 2014). The simulated EnMAP data were also further (spectrally) resampled into Landsat TM data for both dates. This step guarantees a comparable low-dimensional data set to the simulated EnMAP data – contains similar artefacts derived from data preprocessing or varying view angle effects (of the airborne imagery) and excludes any spectral changes due to phenological differences. Additionally, we created a land-cover map of the region through classification of the TM time series, by means of a support vector machine (SVM) classifier. We defined land-cover classes strongly associated with the habitat guilds of the local bird communities (Leitão, Moreira & Osborne 2010): (i) bare soil, (ii) cereal, (iii) grasslands, (iv) woodlands, (v) shrublands and (vi) water. This classification achieved high classification accuracy (overall accuracy of 91·37%; for more details see Table S2) and can thus be considered a high-quality reference product for use as input in our models. The SVM models were run with the IMAGESVM package (Rabe, van der Linden & Hostert 2010), based on the LIBSVM library (Chang & Lin 2011) and implemented in the EnMAP Box (Rabe *et al.* 2012).

All data were compiled to the 125-m radius circular plot level, equivalent to the grain of the bird sampling data (see Table 1). Plot-based average and standard deviation of each individual spectral band were calculated for all Landsat and EnMAP data. Fractions of cover of each class within each plot were calculated from the land-cover map, as well as the number of different classes and the respective Simpson's richness index (Simpson 1949) in a plot. This was done for each bird sampling location (centred in the exact plot location) and for each image pixel (centred in the mid-pixel coordinate).

DATA ANALYSIS

We ran GDM and SGDM models on all data sets: the low-dimensional single-date Landsat TM and land-cover data, and the high-dimensional Landsat time series and EnMAP hyperspectral data. All models were reduced based on variable contribution significance (P-value <0·05). We used the Bray–Curtis dissimilarity metric on all models and did not use the geographical distance as a predictor. The SCCA penalisation parameter grid search was done in 0·1 steps, in a total of 121 possible parameter pair combinations (11 steps for each penalisation parameter). We extracted as many sparse components as possible (i.e. equals the minimum number of variables from both species and environmental matrices) and used the significant ones as final model input.

For the model validation, we extracted a portion (15 samples) of the data in a stratified random manner, following a sparse k-means clustering approach as implemented in R package sparcl (Witten & Tibshirani 2010). All (GDM and SGDM) models were thus built on 100 samples and validated against the remaining samples. This process was iterated three times and the model performance was assessed in the form of the mean (from the three iterations) coefficient of determination (r^2) between observed and predicted values.

To illustrate the use of SGDM for beta-diversity mapping, we used the model on time-series Landsat data to generate a community transition map. For this purpose, the predicted dissimilarities for all sample pairs were transformed using Non-metric Multi-Dimensional Scaling (NMDS; Kruskal 1964). We extracted three NMDS axes, and the factors of these ordinates were then applied to the predicted dissimilarities between the samples and each image pixel (compiled to plot level). Plotting these axes in the red (R), green (G) and blue (B) channels of a

Table 1. Generalised dissimilarity modelling (GDM) and Sparse Generalised Dissimilarity Modelling (SGDM) model results: number of significant variables used in the GDM models, GDM model performance (r^2), penalisation parameters selected for the species matrix (px) and for the environmental matrix (pz), number of significant sparse canonical components used in the SGDM models (SCCs), number of species considered in the resulting components, number of original variables considered in the resulting components and SGDM model performances. The values under parenthesis refer to the respective accounts before eliminating non-significant variables. In the cases when SGDM resulted in a model performance improvement, these were marked in bold

	GDM		SGDM					
			Penalisation		Sparse canonical correlation analysis results			
Dataset	Variables	Performance (r^2)	px	pz	SCCs	Species	Variables	Performance (r^2)
Low-dimensional data sets								
Land-cover map	5 (8)	15·6	0·7	0·5	3 (8)	42 (42)	7 (8)	**18·0**
Landsat TM January	8 (12)	17·9	0·3	0·5	7 (12)	20 (25)	12 (12)	15·4
Landsat TM March	6 (12)	7·1	0·2	0·8	4 (12)	10 (22)	12 (12)	**8·0**
Landsat TM May	6 (12)	15·1	0·9	0·5	4 (12)	42 (42)	12 (12)	10·0
Landsat TM June	5 (12)	7·2	0·7	1·0	5 (12)	42 (42)	12 (12)	**12·1**
Landsat TM July	4 (12)	9·1	0·7	0·4	4 (12)	42 (42)	8 (12)	**10·3**
Landsat TM September	4 (12)	8·6	0·3	0·0	5 (12)	18 (19)	5 (5)	5·7
Landsat TMsim April	6 (12)	6·5	0·2	0·5	6 (12)	13 (19)	12 (12)	**7·5**
Landsat TMsim August	3 (12)	6·4	0·2	0·0	3 (12)	4 (12)	3 (7)	5·5
High-dimensional data sets								
Landsat TM time series	28 (72)	18·8	0·8	0·9	14 (42)	42 (42)	72 (72)	**20·1**
EnMAPsim April	215 (292)	8·9	0·8	0·4	21 (42)	42 (42)	292 (292)	**11·0**
EnMAPsim August	239 (292)	6·4	0·9	0·4	23 (42)	42 (42)	292 (292)	**10·6**

colour image results in a map which illustrates the main community transitions in the study region, where colour changes represent the level of dissimilarity in bird assemblages.

Results

When using low-dimensional data sets, such as single-date multispectral data or land-cover information, the SGDM approach was not consistently successful in improving model performances when compared with GDM. On the other hand, when applied on high-dimensional data sets, the SGDM approach always outperformed the GDM, with model improvements as high as 66% of the original performance (Table 1).

The direct use of remotely sensed spectral (reflectance) data in the models was advantageous in comparison with the use of land-cover data derived from the same data, with a mean performance improvement of 21% on GDM models and 12% on SGDM models. Indeed, the continuous nature of these data closely follows the gradual changes in natural ecosystems over space and time and thus is highly suitable for describing spatial ecological patterns (Foody 1992).

The performance of the single-date models (on multispectral Landsat TM data) varied throughout the different time periods on both methods. The use of SGDM on these data sometimes (but not consistently) resulted in model performance improvements.

Models built on time-series data were always better performing than those built on single-date imagery. Observed model improvements ranged from 5% to 166% for GDM models and from 30% to 256% for SGDM models (depending on the date). The use of the SGDM approach on the time-series data resulted in an improvement of 7% in model performance, when compared with the respective GDM models.

The availability of higher spectral information, using hyperspectral instead of multispectral data, was shown to be advantageous for describing the observed bird communities. Model improvements when using these data were up to 37% with GDM and 92% with SGDM. The SGDM models on hyperspectral data for both dates consistently improved performance in relation to the GDM models, with improvement of up to 66%.

Moderate to low levels of shrinkage on the SCCA (from 0·4 to 1) seemed to be able to deliver good improvements in model performance in comparison with the respective GDMs. This was particularly the case for models run on high-dimensional data, for example simulated EnMAP data for August, with selected penalisation parameters of 0·9 on the species matrix and 0·4 on the environmental matrix. This penalisation still resulted in the use of information from all 42 species and 292 spectral variables in the calculation of the (42) sparse components extracted. The significance test further reduced these into 23 components, however, containing information on all available species and environmental (spectral) variables.

In the predicted community transition map (Fig. 3), the three first NMDS axes represent the main species turnover patterns. A close inspection of the data samples against the ordination map allows the interpretation of the observed species turnover in the region. Indeed, areas with high values in the first axis, that is the red channel (represented in the map in red, pink and yellow colours), are typical pseudo-steppe areas, with

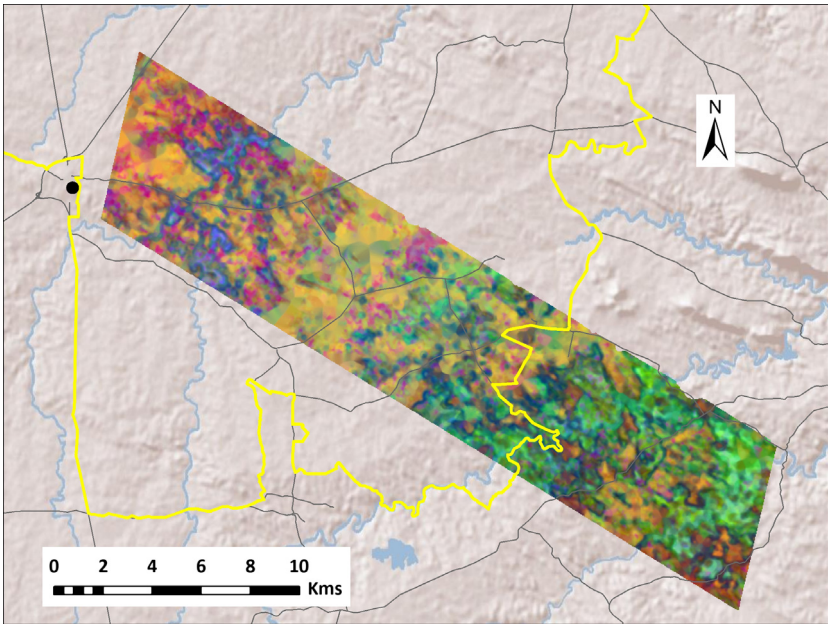

Fig. 3. Example of species compositional turnover mapping in the study area with Sparse Generalised Dissimilarity Modelling based on Landsat time series. The predicted dissimilarities between the sample plots were transformed with Non-metric Multi-Dimensional Scaling. The resulting three axes were applied to the image and visualised on the red, green and blue channels. Roads are represented by the grey lines, the limits of the Castro Verde Special Protection Area by the yellow line and the Castro Verde town by black circle.

the occurrence of species such as little bustard *Tetrax tetrax* or calandra lark *Melanocopypha calandra*. High values in the second NMDS axis (displayed in the green channel) represents areas suitable for species adapted to Mediterranean shrub environments, such as red-legged partridge *Alectoris rufa*, sardinian warbler *Sylvia melanocephala* or Dartford warbler *Sylvia undata*. High values in the third axis (blue channel) represent areas suitable for birds more adapted to fragmenting elements in the steppe mosaic, such as riparian galleries, holm oak woodlands or small farm gardens, such as Iberian azure-winged magpie *Cyanopica cooki* or stonechat *Saxicola torquata*.

The predicted community transition map agrees well with the expected spatial patterns, enabling a meaningful ecological interpretation. For example, we observed the presence of the steppe bird community mainly within the borders of the SPA of Castro Verde as opposed to the dominance of a shrub bird community outside where land abandonment prevails and encroachment is aggravated. By adding new knowledge on the detailed patterns of the community transitions in the study region, this example serves well to illustrate the usefulness of the SGDM for modelling and mapping beta diversity with high-dimensional data.

Discussion

Global environmental change is ongoing, leading to dramatic biodiversity reduction and disturbances in ecological balance with impacts on ecosystem functioning and the provision of ecosystem services (Cardinale *et al.* 2012). Existing and forthcoming new generation global monitoring Earth observation satellites will provide large amounts of high temporally and spectrally resolved data, thus describing the Earth's surface with unprecedented detail. The full depth of these data, such as time series of multispectral or hyperspectral data, although potentially containing suitable information for describing the spatial patterns of beta diversity over large areas, poses challenges for analyses due to their high-dimensional nature.

In this study, we propose a methodological approach which improves the use of high-dimensional (remote sensing) data for modelling biotic communities dissimilarity and turnover via GDM. The Sparse Generalised Dissimilarity Modelling approach (or SGDM) consists of transforming and thus reducing the high-dimensional environmental data through a SCCA (using the species data as ordination constraint), before fitting them with GDM. In this approach, the Lasso-based SCCA (suited for high-dimensional data reduction) (Witten, Tibshirani & Hastie 2009) is parameterised in order to optimise the subsequent GDM performance (in-built in the parameter grid search). The underlying principle of the method is that as the ordination of the environmental data is constrained by the species matrix, the resulting components are associated with the variability (i.e. turnover) in the community, thus making them suitable for modelling its dissimilarity in GDM.

When run on high-dimensional data sets such as a time series of Landsat TM data or simulated EnMAP hyperspectral data, the SGDM consistently outperformed the classical GDM on the same data. In these cases, while there were data reduction through the SCCA ordination (e.g. 72 time-series variables were reduced into 42 sparse canonical components), the extracted components effectively compiled information from all original spectral variables. This was also observed in the cases of the extreme high-dimensional hyperspectral data sets, on which the greater dimension reduction (from 292 variables to 42 components) was translated into greater penalisation of the environmental matrix (lower L_1 bound, in the case 0·4 for both hyperspectral data sets instead of 0·9 for the time-series data), while still keeping information from all original variables. This remained so even after the exclusion of the non-significant variables in the GDM.

As the ordination is used to extract meaningful information from the environmental matrix which is capable of describing the community dissimilarity patterns, high levels of penalisation on the species matrix (determining the down-weighting and potential exclusion of some species) should be avoided in order to assure a strong association between the transformed environmental data and the (full) community data. Indeed, the selected parameters on these (high-dimensional data) models ranged from 0·8 to 0·9 reflecting low penalisation levels. The current code implementation assumes a regular grid of parameter values for both matrices, although this could be adapted in order to for example restrict extreme low L_1 values (high penalisation) on the species matrix.

When run on low-dimensional data sets, for which the SCCA is not well suited, the method showed very ambiguous results, with model performance improvements of up to 69% but also decreases in performance of up to 35%, depending on the data used. Also, the selected penalisation parameters varied from extremely high to extremely low (e.g. from 0·0 to 1·0 on the environmental data) and with no clear association between these and the resulting model performances. We thus consider the SGDM method as unsuitable for these cases.

While Lasso penalisation does not correct for heteroscedasticity (Jia, Rohe & Yu 2013), potentially resulting in sensitivity to high variance species in the SCCA, our tests showed that the SGDM is able to cope well with count data and delivers better results than GDM (for high-dimensional data). However, the usage of the method under extreme heteroscedasticity could result in weaker model performances. Also, although GDM allows the input of presence/absence dissimilarity measures, the applicability of the SGDM approach on occurrence data remains untested.

We thus conclude that SGDM is suitable for use as an alternative to GDM for high-dimensional environmental data sets (e.g. when the number of environmental variables exceeds the number of species), such as time series or high spectrally resolved remote sensing data. Furthermore, SGDM may be applied on repeatedly acquired (remote sensing) data to monitor (through prediction) changes in biodiversity in almost real-time.

Acknowledgements

This research is part of the EnMAP Core Science Team (ECST), which was funded by the German Aerospace Centre (DLR) – Project Management Agency, granted by the Ministry of Economics and Technology (BMWi; grant no. 50EE0949). This work was also partly funded by the European Facility for Airborne Research (EUFAR) in the frame of the HyMedEcos-Gradients project (ref. EUFAR 11-04) and with support from the Airborne Research and Survey Facility (ARSF), Geophysical Equipment Facility (GEF) and Field Spectroscopy Facility (FSF) of the UK's Natural Environmental Research Council (NERC). IC was partially funded by a Portuguese postdoctoral grant from Fundação para a Ciência e Tecnologia (SFRH/BPD/76514/2011). During the field campaigns in Portugal, we generously received support by the Liga para a Proteção da Natureza (LPN). Cornelius Senf and Benjamin Jakimow supported our programming efforts. Andreas Rabe provided valuable input on the use of the EnMAP Box. We also thank two anonymous reviewers and one associate editor for their comments and suggestions that contributed to improve the previous version of the manuscript.

Data accessibility

All R scripts are available as supporting information. The data used in this article are publicly available online from the Dryad Digital Repository (Leitão *et al.* 2015).

References

Baldeck, C.A. & Asner, G.P. (2013) Estimating vegetation beta diversity from airborne imaging spectroscopy and unsupervised clustering. *Remote Sensing*, **5**, 2057–2071.

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

Cardinale, B.J., Duffy, J.E., Gonzalez, A., Hooper, D.U., Perrings, C., Venail, P. *et al.* (2012) Biodiversity loss and its impact on humanity. *Nature*, **486**, 59–67.

Chang, C.-C. & Lin, C.-J. (2011) LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, **2**, 1–27.

Cord, A.F., Meentemeyer, R.K., Leitão, P.J. & Václavík, T. (2013) Modelling species distributions with remote sensing data: bridging disciplinary perspectives. *Journal of Biogeography*, **40**, 2226–2227.

Cord, A.F., Klein, D., Mora, F. & Dech, S. (2014) Comparing the suitability of classified land cover data and remote sensing variables for modeling distribution patterns of plants. *Ecological Modelling*, **272**, 129–140.

De Caceres, M., Legendre, P. & He, F. (2013) Dissimilarity measurements and the size structure of ecological communities. *Methods in Ecology and Evolution*, **4**, 1167–1177.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G. *et al.* (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, **36**, 27–46.

Eldridge, D.J., Bowker, M.A., Maestre, F.T., Roger, E., Reynolds, J.F. & Whitford, W.G. (2011) Impacts of shrub encroachment on ecosystem structure and functioning: towards a global synthesis. *Ecology Letters*, **14**, 709–722.

Faith, D.P., Carter, G., Cassis, G., Ferrier, S. & Wilkie, L. (2003) Complementarity, biodiversity viability analysis, and policy-based algorithms for conservation. *Environmental Science & Policy*, **6**, 311–328.

Feilhauer, H. & Schmidtlein, S. (2009) Mapping continuous fields of forest alpha and beta diversity. *Applied Vegetation Science*, **12**, 429–439.

Ferrier, S., Drielsma, M., Manion, G. & Watson, G. (2002) Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. II. Community-level modelling. *Biodiversity and Conservation*, **11**, 2309–2338.

Ferrier, S., Manion, G., Elith, J. & Richardson, K. (2007) Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. *Diversity and Distributions*, **13**, 252–264.

Foody, G.M. (1992) A fuzzy sets approach to the representation of vegetation continua from remotely sensed data: an example from lowland heath. *Photogrammetric Engineering & Remote Sensing*, **58**, 221–225.

Fuller, R.J. & Langslow, D.R. (1984) Estimating numbers of birds by point counts: how long should counts last? *Bird Study*, **31**, 195–202.

Goslee, S. & Urban, D. (2007) The ecodist package for dissimilarity-based analysis of ecological data. *Journal of Statistical Software*, **22**, 1–19.

Griffiths, P., van der Linden, S., Kuemmerle, T. & Hostert, P. (2012) A pixel-based Landsat compositing algorithm for large area land cover mapping. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **6**, 2088–2101.

Guerin, G.R., Biffin, E. & Lowe, A. (2013) Spatial modelling of species turnover identifies climate ecotones, climate change tipping points and vulnerable taxonomic groups. *Ecography*, **36**, 1086–1096.

Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A. *et al.* (2013) High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science*, **342**, 850–853.

Hooper, D.U., Chapin, F.S., Ewel, J.J., Hector, A., Inchausti, P., Lavorel, S. *et al.* (2005) Effects of biodiversity on ecosystem functioning: a consensus of current knowledge. *Ecological Monographs*, **75**, 3–35.

Jia, J., Rohe, K. & Yu, B. (2013) The Lasso under Poisson-like heteroscedasticity. *Statistica Sinica*, **23**, 99–118.

Kennedy, R.E., Andrefouet, S., Cohen, W.B., Gomez, C., Griffiths, P., Hais, M. *et al.* (2014) Bringing an ecological view of change to Landsat-based remote sensing. *Frontiers in Ecology and the Environment*, **12**, 339–346.

Kerr, J.T. & Ostrovsky, M. (2003) From space to species: ecological applications for remote sensing. *Trends in Ecology and Evolution*, **18**, 299–305.

Kruskal, J.B. (1964) Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, **29**, 115–129.

Legendre, P., Borcard, D. & Peres-Neto, P.R. (2005) Analyzing beta diversity: partitioning the spatial variation of community composition data. *Ecological Monographs*, **75**, 435–450.

Legendre, P. & Gallagher, E.D. (2001) Ecologically meaningful transformations for ordination of species data. *Oecologia*, **129**, 271–280.

Leitão, P.J., Moreira, F. & Osborne, P.E. (2010) Breeding habitat selection by steppe birds in Castro Verde: a remote sensing and advanced statistics approach. *Ardeola*, **57**, 93–116.

Leitão, P.J., Moreira, F. & Osborne, P.E. (2011) Effects of geographical data sampling bias on habitat models of species distributions: a case-study with steppe birds in southern Portugal. *International Journal of Geographical Information Science*, **25**, 439–453.

Leitão, P.J., Schwieder, M., Suess, S., Catry, I., Milton, E.J., Moreira, F. *et al.* (2015) Data from: Mapping beta diversity from space: Sparse Generalized Dissimilarity Modelling (SGDM) for analysing high-dimensional data. *Dryad Digital Repository*, doi: 10.5061/dryad.ns7pv.

Marta-Pedroso, C., Domingos, T., Freitas, H. & de Groot, R.S. (2007) Cost–benefit analysis of the Zonal Program of Castro Verde (Portugal): highlighting the trade-off between biodiversity and soil conservation. *Soil & Tillage Research*, **97**, 79–90.

Masek, J.G., Vermote, E.F., Saleous, N.E., Wolfe, R., Hall, F.G., Huemmrich, K.F., Feng, G., Kutler, J. & Teng-Kui, L. (2006) A Landsat surface reflectance dataset for North America. *IEEE Geoscience and Remote Sensing Letters*, **3**, 68–72.

McKnight, M.W., White, P.S., McDonald, R.I., Lamoreux, J.F., Sechrest, W., Ridgely, R.S. & Stuart, S.N. (2007) Putting beta-diversity on the map: broad-scale congruence and coincidence in the extremes. *Plos Biology*, **5**, 2424–2432.

Moreira, F. (1999) Relationships between vegetation structure and breeding bird densities in fallow cereal steppes in Castro Verde, Portugal. *Bird Study*, **46**, 309–318.

Moreira, F. & Russo, D. (2007) Modelling the impact of agricultural abandonment and wildfires on vertebrate diversity in Mediterranean Europe. *Landscape Ecology*, **22**, 1461–1476.

Moreira, F., Leitão, P.J., Morgado, R., Alcazar, R., Cardoso, A., Carrapato, C. *et al.* (2007) Spatial distribution patterns, habitat correlates and population estimates of steppe birds in Castro Verde. *Airo*, **17**, 5–30.

Moreira, F., Silva, J.P., Estanque, B., Palmeirim, J.M., Lecoq, M., Pinto, M. *et al.* (2012) Mosaic-level inference of the impact of land cover changes in agricultural landscapes on biodiversity: a case-study with a threatened grassland bird. *PLoS One*, **7**, e38876.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B. *et al.* (2012) vegan: Community Ecology Package. R package version 2.0-10. Available at: http://CRAN.R-project.org/package = vegan.

Osborne, P.E., Alonso, J.C. & Bryant, R.G. (2001) Modelling landscape-scale habitat use using GIS and remote sensing: a case study with great bustards. *Journal of Applied Ecology*, **38**, 458–471.

Parviainen, M., Zimmermann, N.E., Heikkinen, R.K. & Luoto, M. (2013) Using unclassified continuous remote sensing data to improve distribution models of red-listed plant species. *Biodiversity and Conservation*, **22**, 1731–1754.

R Development Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabe, A., van der Linden, S. & Hostert, P. (2010) imageSVM, Version 2.1. Available at: www.imagesvm.net.

Rabe, A., B., J., van der Linden, S. & Hostert, P. (2012) EnMAP-Box, Version 1.4. Available at: www.enmap.org.

Reineking, B. & Schröder, B. (2006) Constrain to perform: regularization of habitat models. *Ecological Modelling*, **193**, 675–690.

Rocchini, D. (2007) Distance decay in spectral space in analysing ecosystem beta-diversity. *International Journal of Remote Sensing*, **22**, 2635–2644.

Rocchini, D., Balkenhol, N., Carter, G.A., Foody, G.M., Gillespie, T.W., He, K.S. *et al.* (2010) Remotely sensed spectral heterogeneity as a proxy of species diversity: recent advances and open challenges. *Ecological Informatics*, **5**, 318–329.

Schwieder, M., Leitão, P.J., Suess, S., Senf, C. & Hostert, P. (2014) Estimating fractional shrub cover using simulated EnMAP data: a comparison of three machine learning regression techniques. *Remote Sensing*, **6**, 3427–3445.

Segl, K., Guanter, L., Rogass, C., Kuester, T., Roessner, S., Kaufmann, H., Sang, B., Mogulsky, V. & Hofer, S. (2012) EeteS-The EnMAP End-to-End Simulation Tool. *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **5**, 522–530.

Simpson, E.H. (1949) Measurement of diversity. *Nature*, **163**, 688.

Stuffler, T., Kaufmann, C., Hofer, S., Förster, K.P., Schreier, G., Mueller, A. *et al.* (2007) The EnMAP hyperspectral imager - an advanced optical payload for future applications in Earth observation programmes. *Acta Astronautica*, **61**, 115–120.

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

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & Knight, K. (2005) Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B*, **67**, 91–108.

Tuomisto, H. (2010) A diversity of beta diversities: straightening up a concept gone awry. Part 2. Quantifying beta diversity and related phenomena. *Ecography*, **33**, 23–45.

Turner, W., Spector, S., Gardiner, N., Fladeland, M., Sterling, E. & Steininger, M. (2003) Remote sensing for biodiversity science and conservation. *Trends in Ecology and Evolution*, **18**, 306–314.

United States Geological Survey (2013) USGS Global Visualization Viewer. Available at: http://glovis.usgs.gov/.

Whittaker, R.H. (1960) Vegetation of the Siskiyou Mountains, Oregon and California. *Ecological Monographs*, **30**, 280–338.

Witten, D.M. & Tibshirani, R. (2009) Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, **8**, Article 28.

Witten, D.M. & Tibshirani, R. (2010) A framework for feature selection in clustering. *Journal of the American Statistical Association*, **105**, 713–726.

Witten, D.M., Tibshirani, R. & Hastie, T. (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, **10**, 515–534.

Received 15 December 2014; accepted 17 March 2015
Handling Editor: David Warton

Supporting Information

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

Data S1. Script code and respective description of the functions used within the SGDM approach.

Table S1. Bird community data: species considered in the analysis, their respective frequency of observation (number of plots observed) and abundance (mean number of birds per plot).

Table S2. Accuracy assessment of the land cover classification using Support Vector Machines on the Landsat TM time-stack.

Data S2. SGDM tools.

