Article:
Gavish, Y, O'Connell, J and Benton, TG (2018) Quantifying and modelling decay in forecast proficiency indicates the limits of transferability in land-cover classification. Methods in Ecology and Evolution, 9 (2). pp. 235-244. ISSN: 2041-210X

https://doi.org/10.1111/2041-210X.12870

DR. YONI GAVISH (Orcid ID : 0000-0002-6025-5668)

Article type : Research Article

Handling editor: Dr. Ryan Chisholm

Title:

Quantifying and modelling decay in forecast proficiency indicates the limits of transferability in land-cover classification

Running Title:

Distance-Decay of Forecast Proficiency

Authors:

Gavish, Yoni[1,*] gavishyoni@gmail.com

O'Connell, Jerome[2] oconnell.jerome@gmail.com

Benton, Tim G.[1] T.G.Benton@leeds.ac.uk

[1] School of Biology, Faculty of Biological Sciences, University of Leeds, Leeds, LS2 9JT, UK

[2] School of Biosystems and Food Engineering, University College Dublin, Dublin, D04 N2E5, Ireland

[*] Corresponding author: Gavish Yoni; School of Biology, Faculty of Biological Sciences, University of Leeds, Leeds, LS2 9JT, UK; gavishyoni@gmail.com ; +447599991988.

Type of Paper:

Standard article

Data archiving statement:

Data deposited in the Dryad repository: http://dx.doi.org/10.5061/dryad.s156b. The data package contains the raw data at the objects level, as well as all distance measures between sites and additional meta-data. In addition, the data package contains all the R code needed to reproduce the results. The raw LandMap images are tied under a CEDA academic licences agreement (http://licences.ceda.ac.uk/image/data_access_condition/landmap.pdf) and thus could not be deposited. For accessing the raw images contact: support@ceda.ac.uk. All the details on the dataset are in: http://catalogue.ceda.ac.uk/uuid/51c4273d47ef0130c422eafb2f99d4fe. The information regarding date, X and Y centroid and UK 100 km grid reference supplied in SI1 is required to identify the relevant tiles.

ABSTRACT

1. The ability to provide reliable projections for the current and future distribution of land-cover is fundamental if we wish to protect and manage our diminishing natural resources. Two inter-related revolutions make map productions feasible at unprecedented resolutions - the availability of high-resolution remotely-sensed data and the development of machine-learning algorithms. However, ground-truthed data needed for training models is in most cases spatially and temporally clustered. Therefore, map production requires extrapolation of models from one place to another and the uncertainty cost of such extrapolation is rarely explored. In other words, the focus has mainly been on projections, and less on quantifying their reliability.
2. Using the concept of 'forecast horizon', we suggest that the predictability of land-cover classification models should be quantitatively explored as a continuum against distances measured along multiple dimensions – space, time, environmental and spectral. Focusing on ten agricultural sites from England and using models specifically designed to predict multivariate decay-curves, we ask: how does a model's predictive performance decay with distance? More specifically, we explored if we could predict the proficiency (kappa

statistics) of a model trained in one site when making predictions in another site based on the spatial, temporal, spectral and environmental distances between sites.

3. We found that model proficiency decays with distance between sites in each dimension. More importantly, we found for the first time, that it is possible to predict the performance a model transferred to or from a novel site will have, based on its distances from known sites. The spatial distance variables were the most important when predicting model transferability.
4. Exploring model transferability as a continuum may have multiple usages including predicting uncertainty values in space and time, prioritization of strategies for ground-truth data collection, and optimizing model characteristics for defined tasks.

KEY WORDS:

Signature extension, habitat mapping, remote-sensing, earth-observation, random forest, species-distribution models, community similarity, uncertainty, forecast horizon, predictive ecology

INTRODUCTION

The ability to provide reliable predictions for the current and future distribution of land-covers, habitats, species and communities is fundamental if we wish to protect and manage our diminishing natural resources. Thus, in recent years there has been a surge of interest in developing the scientific backbone of predictive ecology. Among others, predictive ecology is explored in relation to the type of predictions (Mouquet *et al.* 2015), the type of models (Evans *et al.* 2013; Urban *et al.* 2016), model complexity (Wenger & Olden 2012; Evans *et al.* 2013) and the various potential sources of uncertainty (Petchey *et al.* 2015; Cabral *et al.*

2016; Alexander *et al.* 2017). However, we are still far from understanding the limits by which we can predict, project and forecast various quantities over various dimensions.

To better understand predictability in space and time, Petchey *et al.* (2015) suggested the 'ecological forecast horizon' as a unifying concept and defined several terms to facilitate communication. First, 'forecast proficiency' (FP) defines the quantitative measure by which we assess the performance of the model in terms of accuracy and precision. Second, 'forecast proficiency threshold' is the value of the FP above which the predictions are still useful and below which they are not, moving from a qualitative description of models as either 'good' or 'bad' towards task-oriented decisions. Finally, 'forecast horizon' describes the distance along a certain dimension (e.g., temporal, spatial) at which FP falls below the FP threshold. That is, we should explore predictability using quantitative methods as continuous decay curves against distances measured along multiple dimensions. However, such an approach requires knowing the true state of the system at the projected set of conditions. Indeed, Petchey *et al.* (2015) claim that forecast horizon can 'be applied in any situation where the value of a variable is predicted, and there is knowledge about the known or assumed true value of that variable'. Here, we follow the 'gold standard' suggested by Pennekamp *et al.* (2016) and ask whether FP is predictable even in the lack of knowledge of the true value and when FP is likely to simultaneously decay along multiple, potentially interacting distance dimensions.

As competition for land increases on a global basis, there is a growing need for habitat/land-cover (H/LC) maps to guide land-use planning: where does habitat need to be protected, and where can it be developed? At large spatial scales, valuable habitat is often "spared" through protected area status, but at finer spatial scales, land for "sparing" vs land for "sharing" requires knowing in detail its characteristics (Shackelford *et al.* 2015), including ecosystem services provisioning (e.g., Koschke *et al.* 2012), natural capital assessment (Brown *et al.* 2016) and species' distributions (Thuiller, Araujo & Lavorel 2004).

At even finer spatial scales, habitat patches – e.g. hedgerows – may need to be mapped to ensure they can be protected under various laws (EU 2007) due to their role in maintaining landscape connectivity. Finally, reliable H/LC maps are essential for projections of land-cover/land-use change under various scenarios, since inaccuracies in the base map may affect models' outcomes (Alexander *et al.* 2017). Given the increasing need for reliable H/LC models, our aim here is to explore the distance decay of their FP.

In recent years, two developments have been revolutionising the production of H/LC maps. First, open access to remote-sensing data facilitates the production of thematic and high spatial resolution maps over wide extents (Xie, Sha & Yu 2008). Second, the increase in computational power allows machine-learning algorithms to produce reliable rule-sets that can classify H/LC for increasingly large datasets (Gong *et al.* 2013). The limiting factor is now arguably the availability of the ground-truthed data required to train and validate the models (Knorn *et al.* 2009; Heipke 2010). Thus, to fully realize the potential of the data revolution need to optimize the usage of the limited and expensive ground-truthed data.

If the limiting factor in H/LC modelling is the ground-truthing data, it raises a key question that relates directly to the concept of the forecast horizon: to what extent can models trained on data from one location be applied outside its current training extent? To address this question requires data from a new site. But if new data is not available, the assumption is often made that models will perform similarly in new sites to the original one. For H/LC classification models, this assumption was often found to be false (Pax-Lenney *et al.* 2001; Knorn *et al.* 2009). For example, Olthof, Butson and Fraser (2005) found that model proficiency when the inter-sites distance was between 1500-2000 km was 50% of the proficiency when the distance was smaller than 500 km. If indeed transferability of models carries a FP cost, then we regularly over-estimate the accuracy of H/LC maps because ground-truthed data is typically spatially clumped. In fact, Brenning (2012) suggested

integrating model transferability into FP assessment as a way to control for spatial autocorrelation between the training and validation datasets.

Model transferability has also been explored for species distribution models (SDMs). SDM transferability studies can shed light on transferability of H/LC models since the two rely on similar machine-learning algorithms and since SDM can be regarded as a classification problem with 2 classes. Similar emphasis on spatially-structured cross validation was suggested for SDMs as well as H/LC models (Wenger & Olden 2012). Furthermore, multiple studies of the transferability of SDMs found reduced FP of models transferred spatially (Randin *et al.* 2006; Wenger & Olden 2012; Wang & Jackson 2014) or temporally (Tuanmu *et al.* 2011; Huang, Frimpong & Orth 2016). For SDMs, transferability studies mainly focused on the trade-off between model complexity and transferability. In particular, whether more complex models (e.g., Random Forest and Artificial Neutral Networks) which can identify complex (but true) rules that boost the FP at the source site, produce classification errors when transferred because of the specifics of the training site. SDM transferability studies therefore mainly focused on the choice of algorithm, on the number of training variables, and on the size of the training set, treating transferability as a property of the model/algorithm.

Whilst the choice of model clearly affects model transferability, we suggest that transferability should predominantly depend on the similarity between the source site (where the model is trained) and the target site (where the model is predicting). If the source and target site are different from one another, transferability should be low, while if the two sites are similar, transferability should be relatively high. To our knowledge, no SDM studies have explicitly considered the effect of the distance between sites on model transferability and only a handful of papers have done so for H/LC classification.

Here we quantitatively examine the decay of FP with the spatial, temporal, spectral and environmental distance between sites. Specifically, we ask the following questions:

1. How does FP decay with spatial, temporal, spectral and environmental distances and what is the relative importance of these dimensions?
2. Can we predict the FP of models in novel sites using known decay curves?

We focus on the transferability of H/LC models in ten agricultural sites located in central and eastern England, UK. Providing high spatial resolution maps within UK agricultural landscape is of great importance with respect to habitat fragmentation, diversity and sustainability (Benton, Vickery & Wilson 2003; O'Connell, Bradter & Benton 2015). The coverage and spatial distribution of land-cover classes in intensively managed agricultural landscapes can have direct and indirect consequences on ecosystem service delivery and sustainability (Benton 2007). As traditional surveying at regional scales is impractical, remote-sensing based classification models is the only feasible option to delineate such land-covers at the required extent.

METHODS:

Study sites

Ten sites were selected in the UK (Fig. 1a) based on access to ground validation data from previous studies (Relu 2011; O'Connell, Bradter & Benton 2015). Sites varied in size from 1197 - 15136 ha and were located in intensively-managed agricultural landscapes which were primarily dominated by arable crops and permanent grassland (Fig. 1c). The cropped areas were further divided into un-ripened cereal crops, ripened cereal crops, grass and scrub. The non-cropped habitat included trees, hedges, and field margins. We further classified two

additional classes, sparse (fallow or sealed surfaces) and shadow, for a total of nine land-cover classes (Fig 1b).

Data

Colour Infra Red (CIR) aerial photography (CEDA 2014) was acquired for each site (acquisition dates: 2006 to 2010 ; spatial resolution: 0.5m). Data pre-processing was in accordance with O'Connell, Bradter and Benton (2013). Once pre-processed, we applied a segmentation procedure in eCognition (Trimble 2013) to spatially aggregate pixels with similar spectral, textural and geometric properties into objects. A total of 66 spectral variables (Table S1) were created for each object along with 31 environmental variables covering soil maps, topographic properties and climatic conditions. The ground-truth data of the nine land-cover types covered 6.5% of the total population of objects and was assessed using a combination of auxiliary data from previous studies and visual interpretation. A full description of the data preparation is given in Appendix S1 in Supporting Information while Table S1 provide a full list of object-scale variables.

Local H/LC classification models

We used Random Forest (RF, Breiman 2001) for the H/LC classification models. RF takes as input a set of training objects, representing a predefined and finite set of classes, along with relevant explanatory variables. The algorithm then 'learns' the rules by which the explanatory variables distinguish between the different classes. We refer to the site from which the training objects are taken as the source site (S_k). Once a model is trained, it can be used to predict the classes in any other site, referred here as the target site (T_j), such that a full process of training and predicting can be listed as S_kT_j. After predicting, FP is assessed by various proficiency indices that compare the predicted class to the training data (when the

source and target sites are identical, i.e., S_kT_k) or to external validation data (when the source and target sites differ).

For each site we trained a RF model using 1,000 classification trees with the package *randomForest* in *R* (R Core Team 2016), with 72 object-level variables as explanatory variables (Table S1). We then estimated FP within each site (i.e. the 10 S_kT_k pairs), using unweighted kappa statistic (Appendix S1). Next, we used the local model to predict all the objects in all other sites, and quantified kappa separately in each target site. This yielded a total of 90 observed FP for models trained in source site k and applied in target site j (all S_kT_j where $k \neq j$). Since model transferability is not necessarily symmetrical, FP of S_kT_j may differ from the FP of S_jT_k (Randin *et al.* 2006). We explored the symmetry property by calculating the Pearson correlation coefficient between S_kT_j and S_jT_k (for all $k \neq j$ pairs). We calculated the coefficient 500 times, each time randomly allocating the sites of each pair to either k or j. We repeated all RF models 30 times and took the mean kappa for the symmetry analysis.

Distance measures between sites

For each pair of sites, we calculated various measures of the spatial, temporal, spectral and environmental distances. We calculated spatial distance as the Euclidian distance between the sites' centroids. Temporal distance was first calculated as the total number of days separating the dates in which the sites were ground-truthed (hereafter, the 'absolute-days-apart'). However, since two images taken in the same season but different years may be more likely to resemble each other than two images taken at different seasons in the same year, we calculated the days apart as the minimum time between images in Julian days, ignoring years (the 'seasonal shift').

As mentioned above, spectral and environmental variables were collected at the object scale. However, to assess model transferability they needed to be aggregated at the site-scale

and translated to distance measures. Since not all variables can be aggregated using a single methodology we utilized five different complementary methods, each suitable for different types of object-level variables and each capturing a different property:

a) Bray-Curtis dissimilarity for categorical data (e.g., soil-type cover).
b) Differences in the mean values of additive variables such as Normalized Difference Vegetation Index (NDVI).
c) Difference in within-site heterogeneity of object-level variables (e.g., slope).
d) Differences in within-sites change of variable values within a predefined Euclidian distance range (done separately for 0-100 m, 300-400 m and 800-1000 m distance ranges)
e) The Gower dissimilarity (Podani 1999) between two objects randomly selected in two sites, either non-stratified or stratified according to Natural England's provisional agricultural land classification (Natural England 2013) .

Detailed descriptions of methods a-e are given in Appendix S2. In total, we created 68 distance measures between each pair of sites (1 spatial, 2 temporal, 27 spectral and 38 environmental, Table S2). The distance measures allowed exploring model transferability as a continuum with S_kT_k pairs taking a distance of zero (with the exception of mean Gower dissimilarity).

Modelling the distance-decay curves

We expect kappa rates to decay non-linearly with the distance measures. Kappa is also bounded between -1 and 1 (with 1 representing full agreement) and thus may be asymptotic with distance. In many ways kappa therefore resembles community similarity indices, and so we modelled the decay curves using Generalized Dissimilarity Models (GDM, Ferrier *et al.* 2007), an extension of matrix regression that accounts for curvilinear relationships between

response and explanatory variables and for the bounded and asymptotic nature of dissimilarity indices. Since GDMs were designed to model dissimilarity values bounded between 0 and 1 (with 1 representing full disagreement), we transformed kappa prior to analysis: $Kappa_{trans} = (1-Kappa_{orig})/2$ and back-transformed after predicting. To measure the goodness of fit of the GDM model, we used the percent of null deviance explained (Ferrier *et al.* 2007). We kept the default settings of the package *gdm* in *R* (R Core Team 2016) for all GDM analyses.

Question 1: The shape of the decay curves

To explore the general shape of the kappa decay curves we first fitted a full GDM model, with the 68 distance measures. Note that for a given *k* and *j*, the 68 distance measures are identical, regardless of which site is used as source and which as target. However, each pair of sites was included once as S_kT_j and once as S_jT_k due to the potential non-symmetry of observed kappa. We also included the 10 S_kT_k pairs for a total of 100 pairs of sites. We repeated the GDMs separately for each of the 30 RF runs. We fitted a linear mixed effects model of the observed vs. the mean predicted kappa (as a fixed factor) and the IDs of the source and target sites as random factors and kept the marginal R^2 as suggested by Nakagawa and Schielzeth (2013). We explored the change in kappa with all distance measures that were retained in at least 4 out of 30 GDMs.

In addition to the full model, we fitted 14 additional GDM models, in which we used a subset of the distance measures. 4 of the 14 GDMs included distances measures only along one dimension (e.g., spatial), 6 along two dimensions (e.g., spatial and temporal), and 4 along three dimensions (e.g., everything but spectral). Thus, the 15 model generates a nested design, which we used to explore the relative effect of spatial, temporal, spectral and environmental distances on the decay curve (Appendix S3). We repeated all GDM analyses

separately for each of the 30 RF runs and ranked the GDM models based on the mean percent deviance explained over all 30 runs.

Question 2: Predicting model proficiency in novel sites

To explore if the decay curves can predict the FP of models in novel sites, for each S_kT_j we needed to compare the predicted kappa to the observed, when at least one of the sites is set aside and not included in the GDM analysis. Thus we can further distinguish between two sub-questions (see Appendix S1 for more details):

- **Source sites are novel:** if *k* is a novel source site, we can explore our ability to predict the FP that a model trained in *k* will have in any other site. This can then be used to prioritise sites for ground-truthing based on their ability to predict in other sites. Here, we are predicting the ability of the novel site to forecast.
- **Target sites are novel:** If *j* is a novel target site, we can explore our ability to predict the FP that a classification model trained in any other site will have in site *j*. This can then be used to map the maximal FP across the entire extent (i.e., uncertainty maps). Here, we are predicting the ability of a site to be forecasted.

To answer these questions focusing on site *z* as the novel site, we first removed from the 100 pairs of sites all 19 pairs in which site *z* was either a source, a target, or both. We then fitted a GDM for the remaining 81 pair using all 68 distance measures. Next we used the GDM model to predict the kappa for the 19 pairs. Note that the predicted kappa for S_jT_z and S_zT_j is identical, yet the two questions differ since the observed kappa of each pair is not necessarily identical. We repeated the above procedure for all sites separately for each of the 30 RF runs

and took for each pair the mean predicted kappa over all runs. We fitted a linear mixed effect model of the observed vs. predicted kappa as described above.

RESULTS

The data and R codes required to reproduce our results was submitted to Dryad (Gavish, Benton & O'Connell 2017). We observed a drop in FP (measured by kappa values), whenever a model is transferred from a source site to a different target site. The observed self-kappa had a mean of 0.780 ($\pm$0.033 S.E) and ranged between 0.727-0.821, while the observed transferred-kappa had a mean of 0.360 ($\pm$0.004) and ranged between 0.124-0.676. The transferability of sites was not entirely symmetrical, although a clear trend was observed between S_kT_j and S_jT_K (Fig. 2) with mean Pearson correlation coefficient of 0.67 ± 0.01 (sd).

Question 1: The shape of the decay curves

The GDM analysis that included all 68 distance measures accurately predicted kappa (Fig. 3a, marginal R^2=0.94). Among the 68 distance measures, the GDM identified 26 measures as important in at least 4 of the 30 runs, including the spatial distance, the two temporal distance, 13 spectral measures and 10 environmental measures (Table S2). The clearest decay curves were observed against the spatial distance and the absolute-days-apart, with less clear trend against the seasonal-shift (Fig. 4). For these 3 distance measures we found clear decay curves of class level accuracies for some but not all of the thematic classes (Appendix S4). Class level accuracies did not change in a systematic manner with the number of objects in the ground-truth data of the target site, suggesting that class imbalance had little effect on the kappa decay curves (Appendix S4).

For the spectral distance measures (Fig. 4 and Fig. S1), the GDM models retained various distance measures based on the object-level reflectance in red and Near-Infra-Red (NIR).

This included changes in mean values (method b) and differences in within-site heterogeneity (method c). Interestingly, the objects within each site were created from multiple pixels, and we found that differences in within-object heterogeneity (based on method b) in various variables (red, NIR, Canny edge detection, EVI2, see Table S2) also affected model transferability. Finally, the GDM flagged two of the Gower-based measures (method e) as important: those stratified by Natural England's provisional agricultural land classification grades 2 ("Very-Good") and 4 ("Poor") (Natural England 2013).

The 10 environmental distance measures (Fig. 4 and Fig. S2) included difference between sites in the standard deviations of rainfall in March, June and August, and the number of growing days (method b). In addition, the GDM retained the difference between sites in the relative cover of different soil types (method a). Finally, difference between sites in spatially structured heterogeneity (method d) of various topographic object-level variables (e.g., slope, elevation) were identified as important.

The nested GDM analysis (Fig. 5, Appendix S3) revealed that the GDM that included all 68 distance measures explained the highest percent of null deviance (94.20% $\pm$0.02 S.E). The analyses identified the spectral dimension as the most important, followed closely by the environmental dimension, with the temporal and spatial dimensions lagging behind, though still performing well (73.38%$\pm$0.28 for the temporal-only GDM). These results suggest that GDMs provide a good modelling framework to explore model transferability, and that model transferability is related to the various distance measures.

Question 2: Predicting model proficiency in novel sites

The GDMs fitted when excluding one site explained on average 95.01% ($\pm$0.20 S.E) of the null deviance. There was a strong correlation between the observed and predicted kappa, both when the novel sites where source (Fig. 3b) and target (Fig. 3c). In both cases, the linear

mixed effect model returned significant results for the fixed term with marginal R^2 of 0.68 and 0.61 when the novel sites where source or target, respectively.

DISCUSSION

The ”big data” revolution provides the potential for spatially explicit environmental data to become available in unprecedented ways through remote-sensing and data analytics. Such data can revolutionise our ability to manage our increasingly pressured environment in order to preserve ecosystem services as competition for natural resources intensifies. However, complex models applied to complex remotely sensed data requires detailed ground-truth data for training, which is expensive to collect. It is therefore vital that we understand the consequences of extrapolating such models across landscapes. If model transferability is low, it implies that our ability to use remotely-sensed data, in the absence of ground-truthed data, will also be low, providing a significant constraint on the ability to realise the potential of increased availability of imagery.

In this study, we found that FP decayed with the spatial, temporal, spectral and environmental distance between source and target sites (Fig. 4). Importantly, we also found that the decay curves can be used to predict the proficiency cost when transferring a model from one site to another, even if one of the sites is completely novel (Fig. 3). Finally, we found that among the various sets of distance measures, those based on spectral variables are the most important to predicting model transferability (Fig. 5). Our results provide some positive reassurance: models can provide predictive ability in novel landscapes, away from the place and time where they were trained. But perhaps more importantly, their proficiency is likely to be estimable; allowing for the provision of both spatially-explicit H/LC data and its certainty to be incorporated into land-management decisions, when field survey data may

not be available (Cabral *et al.* 2016). Although we have focused on H/LC classification, we suspect our results will broadly apply for SDMs.

In this paper, we integrated knowledge and methods from two interrelated disciplines: remote-sensing and ecology. From remote-sensing we took the quantitative approach that explores model transferability as a continuum against spatial, temporal or spectral distances (Pax-Lenney *et al.* 2001; Olthof, Butson & Fraser 2005; Knorn *et al.* 2009; Laborte, Maunahan & Hijmans 2010; Verhulp & Van Niekerk 2016). From ecology, we took knowledge on the shape of the decay curves and the statistical tools appropriate for exploring them (Nekola & White 1999; Ferrier *et al.* 2007; Rocchini 2007). Combined, the exploration of model transferability with GDM can prove to be an important contribution to the growing tool-box of predictive ecology.

The shape of the decay curves

Among the various distance measures, the clearest decay curves were observed against the spatial distance and the absolute days apart temporal distance (Fig. 4a,b, respectively). In fact, spatial distance plus the two temporal distance measures were enough to explain around 80% of the null deviance of model transferability (Fig. 5, Appendix S3). The strong effect of spatial and temporal distances on model transferability are in accordance with other works on signature extension (in remote-sensing, the process of increasing the spatial and temporal range over which a model can be used to classify data without significant loss of FP; Olthof, Butson & Fraser 2005; Laborte, Maunahan & Hijmans 2010; Verhulp & Van Niekerk 2016).

Although the decay patterns against the spatial and temporal distances were very clear, we found that model transferability is best predicted by the spectral distance between sites (Fig. 5). We had more environmental distance measures than spectral ones, so it is unlikely to be simply an effect of the number of variables. Equally, although all of the spectral variables and

many of the environmental variables were measured at the object level (unlike the spatial and temporal distances), there was no one-to-one relation between the variable importance in the RF and selection in the GDMs. For example, the object-level reflectance in the Green band, which was identified as the most important object-level variable in most RF models, was only retained by the GDM once (out of 30 runs). Thus, our results suggest that the variables that are important when fitting a model are not necessarily the ones needed to understand its transferability.

Implications to the forecast horizon concept

We have found that model transferability is better predicted when exploring concurrently multiple distance measures from multiple dimensions (Fig. 5). In such a context, it is difficult to consider a single forecast horizon. For example, focusing on Fig. 4a and arbitrarily deciding on a FP threshold of kappa=0.5 suggest a forecast horizon of approximately 50 km. However, the results of 50 km is conditioned upon the distance measured along other dimensions. For example, sites 50 km apart are expected to have kappa > 0.5 if they are less than 90 days apart, but values drop sharply if the temporal distance is larger (Fig. 5b). In that respect, statistical tools such as GDMs can provide predictions for a wide range of combination of distances along multiple dimensions.

In fact, we used here the GDMs to test the 'gold standard' of model predictability (Pennekamp *et al.* 2016) by applying the models to situations different from what they were trained for, along multiple dimensions and by keeping the training and testing data independent. That is, by setting each time a different site aside and predicting its proficiency as a novel target or source we extrapolated the complex proficiency decay along multiple dimensions to novel combinations of conditions. The importance of keeping to this 'gold standard' becomes clear when comparing the predictive abilities of the GDM when the sites

are not set aside and when they are set aside (Fig. 3). If we had not kept to the 'gold standard' we would claim that the predictability of model proficiency is considerably higher than it actually is. Nonetheless, we found that the transferability of models to or from novel sites is somewhat predictable (Fig. 3b,c). While such results are encouraging, we also found considerable decrease in FP with distance (Fig. 4), suggesting that for our case-study, H/LC maps do not possess high levels of transferability. This leads to a primary question: what is the value of a classification model that has relatively low accuracy?

Clearly the answer is context dependent, but being able to create a map and being able to estimate its uncertainty, is likely to be of utility in some situations where decisions need to be made in the absence of better information. Our models focused on land-cover classes that maintain biodiversity connectivity. Whilst a transferred model with low kappa may not be enough to make explicit decisions on conservation priorities in novel landscapes (i.e., ensuring maximum connectivity by protecting specific objects), it may be enough to monitor the overall connectivity of the landscape in relation to other landscapes or as part of a national monitoring system.

Why explore model transferability?

We see several potential usages of exploring transferability in conservation and management of land-covers, habitats and species:

i. *Mapping uncertainty:* we can predict the FP that each known site will have in any novel site. Therefore, we can map classification uncertainty and identify areas for which we currently cannot provide a reliable classification model.

ii. *Uncertainty of future projections:* Future projections based on climate/land-use change scenarios are usually not accompanied with any estimate of proficiency. A GDM

exploring transferability of models trained at different times for a single site may be extrapolated to provide uncertainty levels to future projections.

iii. *Optimize the collection of survey/ground-truthing data:* In habitats and regions throughout the world where extensive ground-truthing over large areas is prohibitive due to financial, legislative or environmental barriers, understanding when predictive performance declines can aid in the design of optimal data-collection, via the selection of sites where data has the greatest utility in prediction.

iv. *Averaging of local models to optimise model performance in novel sites:* utilising the combined predictive performance of local models in novel sites using kappa weighting. It is possible therefore to give higher weights to sites that are likely to provide more accurate classification in the novel site.

v. *Comparing classification algorithms:* much on the work in the SDM literature has focused on the trade-off between model complexity and model transferability. The forecast horizon concept puts this trade-off on a continuum, in which a more complex algorithm starts at higher proficiency but decays fast, while simpler algorithms start with a lower proficiency but decay slower. If different algorithms decay differently, we can select the best algorithm for a given task, or use the predicted performance as weights.

vi. *Variable selection:* Traditionally, variable selection methods prioritise variables that increase local FP. However, some explanatory variables that play little role in maximizing local FP may be fundamental to model transferability. On the other hand, some variables may be very important for the local FP, but may reduce the transferability of the model. Comparing the decay curves with and without certain variables may help identify such variables.

Conclusions

We show that model transferability declines with spatial, temporal, spectral and environmental distance. This potentially means that any classification, which is trained on a subset of data that is not fully representative of an area's variance, may potentially be error prone. For example, maps based on spatially-clumped ground-truth data may have larger errors between the clumps than within them. Whilst this may be difficult to avoid, we show how, in principle, it would be possible to produce spatial maps of prediction uncertainty as well as H/LC classification. Reliable H/LC maps over wide extent at fine resolution is often a pre-requirement of science-based conservation and management. However, ecological field data is increasingly a limiting factor in producing H/LC maps, so it is unavoidable to extrapolate classifications beyond the area of ground-truthed data. Here we demonstrated the predictive ability and therefore the proficiency costs of extrapolating beyond the extent of the original data with respect to land-cover classification. This may assist in making informative decision when extrapolation is needed, through facilitating the development of methodologies and protocols for wide-scale H/LC mapping in conservation and environmental resource management. More specifically, exploring model transferability as a continuum may open new possibilities of mapping uncertainty, optimising site selection and selecting the optimal algorithm for a given task and environment.

REFERENCES

Alexander, P., Prestele, R., Verburg, P.H., Arneth, A., Baranzelli, C., Batista e Silva, F., Brown, C., Butler, A., Calvin, K., Dendoncker, N., Doelman, J.C., Dunford, R., Engström, K., Eitelberg, D., Fujimori, S., Harrison, P.A., Hasegawa, T., Havlik, P., Holzhauer, S., Humpenöder, F., Jacobs-Crisioni, C., Jain, A.K., Krisztin, T., Kyle, P., Lavalle, C., Lenton, T., Liu, J., Meiyappan, P., Popp, A., Powell, T., Sands, R.D., Schaldach, R., Stehfest, E., Steinbuks, J., Tabeau, A., van Meijl, H., Wise, M.A. & Rounsevell, M.D.A. (2017) Assessing uncertainties in land cover projections. *Global Change Biology,* **23,** 767-781.

Benton, T.G. (2007) Ecology - Managing farming's footprint on biodiversity. *Science,* **315,** 341-342.

Benton, T.G., Vickery, J.A. & Wilson, J.D. (2003) Farmland biodiversity: is habitat heterogeneity the key? *Trends in Ecology & Evolution,* **18,** 182-188.

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

Brenning, A. (2012) Spatial cross-validation and bootstrap for the assessment of predcition rules in remote sensing: the R package sperrorest. *2012 Ieee International Geoscience and Remote Sensing Symposium*, pp. 5372-5375.

Brown, C., King, S., Ling, M., Bowles-Newark, N., Ingwall-King, L., Wilson, L., Pietilä, K., Regan, E.C. & Vause, J. (2016) Natural Capital Assessments at the National and Sub-national Level. *UNEP-WCMC, Cambridge, UK*.

Cabral, R.B., Halpern, B.S., Costello, C. & Gaines, S.D. (2016) Unexpected Management Choices When Accounting for Uncertainty in Ecosystem Service Tradeoff Analyses. *Conservation Letters***,** n/a-n/a.

CEDA (2014) Colour InfraRed (CIR) data for England and Wales, in: Analysis, C.f.E.D. (Ed.), Landmap Optical Earth Observation Collection, 1 ed. . *NERC Earth Observation Data Centre*.

EU (2007) Habitats Directive. In: Commission, E. (Ed.), Article 10.

Evans, M.R., Grimm, V., Johst, K., Knuuttila, T., de Langhe, R., Lessells, C.M., Merz, M., O'Malley, M.A., Orzack, S.H., Weisberg, M., Wilkinson, D.J., Wolkenhauer, O. & Benton, T.G. (2013) Do simple models lead to generality in ecology? *Trends in Ecology & Evolution,* **28,** 578-583.

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.

Gavish, Y., Benton, T.G. & O'Connell, J. (2017) Data from: Quantifying and modelling decay in forecast proficiency indicates limits of transferability in land-cover classification (MEE-17-06-563). doi:10.5061/dryad.s156b. *Methods in Ecology and Evolution*.

Gong, P., Wang, J., Yu, L., Zhao, Y.C., Zhao, Y.Y., Liang, L., Niu, Z.G., Huang, X.M., Fu, H.H., Liu, S., Li, C.C., Li, X.Y., Fu, W., Liu, C.X., Xu, Y., Wang, X.Y., Cheng, Q., Hu, L.Y., Yao, W.B., Zhang, H., Zhu, P., Zhao, Z.Y., Zhang, H.Y., Zheng, Y.M., Ji, L.Y., Zhang, Y.W., Chen, H., Yan, A., Guo, J.H., Wang, L., Liu, X.J., Shi, T.T., Zhu, M.H., Chen, Y.L., Yang, G.W., Tang, P., Xu, B., Giri, C., Clinton, N., Zhu, Z.L. & Chen, J. (2013) Finer resolution observation and monitoring of global land cover: first mapping results with Landsat TM and ETM+ data. *International Journal of Remote Sensing,* **34,** 2607-2654.

Heipke, C. (2010) Crowdsourcing geospatial data. *Isprs Journal of Photogrammetry and Remote Sensing,* **65,** 550-557.

Huang, J., Frimpong, E.A. & Orth, D.J. (2016) Temporal transferability of stream fish distribution models: can uncalibrated SDMs predict distribution shifts over time? *Diversity and Distributions,* **22,** 651-662.

Knorn, J., Rabe, A., Radeloff, V.C., Kuemmerle, T., Kozak, J. & Hostert, P. (2009) Land cover mapping of large areas using chain classification of neighboring Landsat satellite images. *Remote Sensing of Environment,* **113,** 957-964.

Koschke, L., Fürst, C., Frank, S. & Makeschin, F. (2012) A multi-criteria approach for an integrated land-cover-based assessment of ecosystem services provision to support landscape planning. *Ecological Indicators,* **21,** 54-66.

Laborte, A.G., Maunahan, A.A. & Hijmans, R.J. (2010) Spectral Signature Generalization and Expansion Can Improve the Accuracy of Satellite Image Classification. *Plos One,* **5,** 9.

Mouquet, N., Lagadeuc, Y., Devictor, V., Doyen, L., Duputié, A., Eveillard, D., Faure, D., Garnier, E., Gimenez, O., Huneman, P., Jabot, F., Jarne, P., Joly, D., Julliard, R., Kéfi, S., Kergoat, G.J., Lavorel, S., Le Gall, L., Meslin, L., Morand, S., Morin, X., Morlon, H., Pinay, G., Pradel, R., Schurr, F.M., Thuiller, W. & Loreau, M. (2015) REVIEW: Predictive ecology in a changing world. *Journal of Applied Ecology,* **52,** 1293-1310.

Nakagawa, S. & Schielzeth, H. (2013) A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution,* **4,** 133-142.

Natural England (2013) Provisional Agricultural Land Classification (ALC).

Nekola, J.C. & White, P.S. (1999) The distance decay of similarity in biogeography and ecology. *Journal of Biogeography,* **26,** 867-878.
O'Connell, J., Bradter, U. & Benton, T.G. (2013) Using high resolution CIR imagery in the classification of non-cropped areas in agricultural landscapes in the UK. *SPIE Remote Sensing. International Society for Optics and Photonics* pp. 1-15. Dresden, Germany.
O'Connell, J., Bradter, U. & Benton, T.G. (2015) Wide-area mapping of small-scale features in agricultural landscapes using airborne remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing,* **109,** 165-177.
Olthof, I., Butson, C. & Fraser, R. (2005) Signature extension through space for northern landcover classification: A comparison of radiometric correction methods. *Remote Sensing of Environment,* **95,** 290-302.
Pax-Lenney, M., Woodcock, C.E., Macomber, S.A., Gopal, S. & Song, C. (2001) Forest mapping with a generalized classifier and Landsat TM data. *Remote Sensing of Environment,* **77,** 241-250.
Pennekamp, F., Adamson, M.W., Petchey, O.L., Poggiale, J.-C., Aguiar, M., Kooi, B.W., Botkin, D.B. & DeAngelis, D.L. (2016) The practice of prediction: What can ecologists learn from applied, ecology-related fields? *Ecological Complexity*.
Petchey, O.L., Pontarp, M., Massie, T.M., Kéfi, S., Ozgul, A., Weilenmann, M., Palamara, G.M., Altermatt, F., Matthews, B., Levine, J.M., Childs, D.Z., McGill, B.J., Schaepman, M.E., Schmid, B., Spaak, P., Beckerman, A.P., Pennekamp, F. & Pearse, I.S. (2015) The ecological forecast horizon, and examples of its uses and determinants. *Ecology letters,* **18,** 597-611.
Podani, J. (1999) Extending Gower's general coefficient of similarity to ordinal characters. *Taxon*.
R Core Team (2016) R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org*.
Randin, C.F., Dirnböck, T., Dullinger, S., Zimmermann, N.E., Zappa, M. & Guisan, A. (2006) Are niche-based species distribution models transferable in space? *Journal of Biogeography,* **33,** 1689-1703.
Relu (2011) Knowledge Portal. In. Rural Exonomy and Land Use Programme.
Rocchini, D. (2007) Distance decay in spectral space in analysing ecosystem beta-diversity. *International Journal of Remote Sensing,* **28,** 2635-2644.
Shackelford, G.E., Steward, P.R., German, R.N., Sait, S.M. & Benton, T.G. (2015) Conservation planning in agricultural landscapes: hotspots of conflict between agriculture and nature. *Diversity and Distributions,* **21,** 357-367.
Thuiller, W., Araujo, M.B. & Lavorel, S. (2004) Do we need land-cover data to model species distributions in Europe? *Journal of Biogeography,* **31,** 353-361.
Trimble (2013) eCognition Developer. In. Trimble Geospatial Imaging Arnulfstrasse 126, 80636 Munich, Germany.
Tuanmu, M.N., Vina, A., Roloff, G.J., Liu, W., Ouyang, Z.Y., Zhang, H.M. & Liu, J.G. (2011) Temporal transferability of wildlife habitat models: implications for habitat monitoring. *Journal of Biogeography,* **38,** 1510-1523.
Urban, M.C., Bocedi, G., Hendry, A.P., Mihoub, J.-B., Pe'er, G., Singer, A., Bridle, J.R., Crozier, L.G., De Meester, L., Godsoe, W., Gonzalez, A., Hellmann, J.J., Holt, R.D., Huth, A., Johst, K., Krug, C.B., Leadley, P.W., Palmer, S.C.F., Pantel, J.H., Schmitz, A., Zollner, P.A. & Travis, J.M.J. (2016) Improving the forecast for biodiversity under climate change. *Science,* **353**.
Verhulp, J. & Van Niekerk, A. (2016) Effect of inter-image spectral variation on land cover separability in heterogeneous areas. *International Journal of Remote Sensing,* **37,** 1639-1657.
Wang, L.F. & Jackson, D.A. (2014) Shaping up model transferability and generality of species distribution modeling for predicting invasions: implications from a study on Bythotrephes longimanus. *Biological Invasions,* **16,** 2079-2103.
Wenger, S.J. & Olden, J.D. (2012) Assessing transferability of ecological models: an underappreciated aspect of statistical validation. *Methods in Ecology and Evolution,* **3,** 260-267.

Xie, Y., Sha, Z. & Yu, M. (2008) Remote sensing imagery in vegetation mapping: a review. *Journal of Plant Ecology,* **1,** 9-23.

SUPPORTING INFORMATION

- **Appendix S1:** Details on image analyses, creation of object level variables, model evaluation and site novelty.
- **Appendix S2:** The five methods used to translate object-level variables to between-sites distance measures
- **Appendix S3:** Nested GDMs to explore the relative importance of spatial, temporal, spectral and environmental distances.
- **Appendix S4:** User's and producer's accuracy decay against the spatial and temporal distances.
- **Table S1:** full list of object-level variables, their description and their usage in various stages of the analysis.
- **Table S2:** the 68 distance measures, the object-level variable they are based on, the method used in their calculation and the number of times they were retained in the GDM.
- **Figure S1**: The change in observed kappa with the 13 spectral distance measures retained by the at least four GDM.
- **Figure S2:** The change in observed kappa with the 10 environmental distance measures retained by the at least four GDM.

FUNDING

YG and TGB were partially financed by the EU BON project (www.eubon.eu) which is a 7th Framework Programme funded by the European Union under Contract No. 308454. JOC was financed by the Nigel Bertram Charitable Trust, UK.

AUTHOR CONTRIBUTION

YG and JO developed the original concept. JO accumulated the ground-truth data and run all GIS procedures. YG analysed the data and wrote the first draft of the paper. All authors contributed considerably to later versions.

Figure legends:

Figure 1: (a) The location of the 10 sites in the UK. (b) the class hierarchy used in this manuscript. (c) enlarged site 10 with colour infra-red imagery (CIR) where pink/red represents areas of high chlorophyll content and green represents areas of sparse vegetation or low chlorophyll content (© GeoPerspectives supplied by Bluesky 2017). (d) enlarged section of c, with training objects (yellow) overlaid.

Figure 2: The symmetry of model transferability. The kappa when training a Random Forest model in site A and predicting in site B (S_AT_B) against the kappa when training a model in site B and predicting in A (S_BT_A). For a given pair of sites two points are drawn, once when site 1 is A and site 2 is B and once when site 2 is A and site 1 is B (linked by a segment). These point are symmetrical around the line of unity (dashed). However, transferability is not fully symmetrical since the points do not fall on the line of unity. The mean Pearson correlation coefficient is the average over 500 random selection of one point from each pair

Figure 3: The observed kappa against the kappa predicted by the GDM model for the 100 pairs of sites when (a) none of the sites are novel, (b) the source sites are novel and (c) the target sites are novel. Solid line and text are the results of a linear mixed effect model. Dashed line is the line of unity.

Figure 4: Examples of the change of observed kappa with (a) spatial, (b,c) temporal, (d,e,f), spectral and (g,h,i) environmental distances between sites. The solid line is a loess curve ($\pm$ confidence intervals). For the spectral and environmental distance, the *Objects* line describes the object-level variable on which the distance measure is based, while the *Sites* line refers to methods a-e in the main text. ALC G2 is the agricultural land classification grades 2. Additional plots are found in Figures S1 and S2.

Figure 5: The percent of null deviance explained by each of the 15 nested GDM models. Bottom panel provides the sets of distance measures included in each GDM. Note the grouping of spectral, environmental, spatial and temporal sets of distance measures. For each GDM, we observed very little variance between the 30 runs (boxplot).