
Henrys, P.A.; Bee, E.J.; Watkins, J.W.; Smith, N.A.; Griffiths, R.I. 2015.
Mapping natural capital: optimising the use of national scale datasets.
Ecography, 38 (6). 10.1111/ecog.00402

Mapping Natural Capital: Optimising the use of national scale datasets

Henrys, P. A.[1], *Bee, E. J.*[2], *Watkins, J. W.*[1], *Smith, N.A.*[3] *and Griffiths, R. I.*[4]

[1] Centre for Ecology and Hydrology, Lancaster Environment Centre, Library Avenue, Bailrigg, Lancaster, LA1 4AP, UK.

[2] British Geological Survey, Environmental Science Centre, Nicker Hill, Keyworth, Nottingham, NG12 5GG, UK.

[3] British Geological Survey, Murchison House, West Mains Road, Edinburgh, EH9 3LA, UK.

[4] Centre for Ecology and Hydrology, Maclean Building, Benson Lane, Crowmarsh Gifford, Wallingford, Oxfordshire, OX10 8BB, UK.

Correspondence: Peter Henrys, Centre for Ecology and Hydrology, Lancaster Environment Centre, Library Avenue, Bailrigg, Lancaster, LA1 4AP, UK.

Email: pehn@ceh.ac.uk

Tel: +44(0)1524 595839

Abstract

Understanding the spatial distribution of specific environmental variables and the interdependencies of these variables is crucial for managing the environment in a sustainable way. Here we discuss two methods of mapping – a Geographical Information System classification-based approach and a statistical model-based approach. If detailed, spatially comprehensive covariate datasets exist to complement the ecological-response data, then using a statistical model-based analysis provides the potential for greater understanding of underlying relationships, as well as the uncertainty in the spatial predictions. Further, the model-based approach facilitates scenario testing. Although similar methods are already adopted in species distribution modeling, the flexibility of the model framework used is rarely exploited to go beyond modeling occupancy or suitability for a single species, into modeling complex derived metrics such as community composition and indicators of natural capital. As an example, we assess the potential benefits of the statistical model-based approach to mapping natural capital through the use of two national survey datasets; The Centre for Ecology and Hydrology (CEH) Land Cover Map (LCM) and the British Geological Survey's (BGS) Parent Material Model (PMM), to predict national soil microbial community distributions based on data from a sample of > 1000 soils covering Great Britain. The results are mapped and compared against a more traditional, land classification-based approach. The comparison shows that, although the maps look broadly similar, the model-based approach provides better overall spatial prediction, and the contribution of individual model terms (along with their uncertainty) are far easier to understand and interpret, whilst also facilitating any scenario testing. We therefore both recommend the use of spatial statistical modelling techniques to map natural capital and anticipate that they will become more prominent over the forthcoming years.

Introduction

The Millennium Ecosystem Assessment (2005) and more recently in the UK, the National Ecosystem Assessment (2011), stress the importance of ecosystems and understanding the interdependencies between their underlying drivers of change (Carpenter et al. 2009; Feld et al. 2009; Norgaard 2009). Ecosystem 'natural capital' can be identified, according to Costanza and Daly (1992), as the "assets" or "stock that yields a flow of valuable goods or services into the future". This concept of "natural capital" and "flow of goods" has gained traction in recent years and has been used as a way of bridging the scientific-economic-policymaking divide, enabling the potential impact of ecosystem modification to be better evaluated, and more meaningfully incorporated, into decisions affecting society (National Research Council, 2005; Millennium Ecosystem Assessment, 2005). Knowledge regarding the spatial distribution of ecological systems and the natural capital stocks that they produce is of crucial importance for managing the effects of human pressure and environmental change on natural resources (Swetnam et al. 2011; Naidoo and Ricketts 2006).

In order to investigate spatial distribution and variation in natural capital, it is crucial to make use of all available data, both on the natural capital indicator itself and on complementary datasets that are *a priori* thought to drive changes in this response – it is important from the outset that ecological understanding of the system and any synthesis of it are clearly thought about (Austin, 2002). This is to provide unbiased estimates of stocks of natural capital and related ecosystems, enabling planners and policy makers to identify the most economically or environmentally desirable trade-offs (Turner et al. 2010; Nengwang et al. 2009). For example, the availability of suitable habitat for wild pollinator populations may vary depending on the relative strength of the different abiotic and biotic environmental drivers present, such as climate, soil, geology or types of habitats. One approach to investigate the spatial distribution of natural capital may be based on a geographical stratification of the region of interest according to environmental conditions. However, a simple environmental stratification or categorical classification does not provide the flexibility to analyse different drivers, measure their relative strength in determining how stocks are currently distributed, or predict how these may change under future management or environmental change scenarios. All of these require a more flexible approach capable of making best use of a range of source data.

Two examples of the Geographical Information System (GIS) classification-based approach illustrate its shortcomings. For example, the US Geological Society (USGS) generated a map of standardised terrestrial ecosystems across the US that could be useful for studies of the production and value of ecosystem goods and services and indicators thereof (Sayre et al. 2009). The map is derived by classifying areas according to a set of environmental covariates that describe features such as climate and geology. The Institute of Terrestrial Ecology's (ITE) land classification of Great Britain provides a similar map of environmental classes, defined according to a clustering technique imposed on a multivariate ordination, and was based on multiple covariate data sets such as geology, topography and climate (Bunce et al. 1996). The assumption made is that all important covariate effects are accounted for in the classification. These classification maps of environmental or ecosystem strata can provide a basis on which one can overlay, and hence map, specific indicators of natural capital based on the spatial pattern of the strata. However, any further inference, uncertainty analysis, or testing of assumptions and hypotheses, is not possible as the classes are fixed and we cannot disaggregate which drivers are most important for understanding the regional variation or extent of the natural capital indicator in question. Furthermore, one can only make inferences regarding change and association within the existing classification structure, and they cannot be used to predict the outcome after environmental changes (such as climate change or different land use regimens). Such GIS classification-based approaches are commonly used to map natural capital and ecosystem service indicators (eg. Norton et al., 2012; Troy and Wilson, 2006; Raymond et al., 2009; Costanza et al., 2006), but any uncertainty analysis or understanding of spatial dependence is rarely explored as the classification approach does not lend itself to this.

In contrast, using spatial statistical models with an ability to compensate for or make use of spatial autocorrelation, the high quality, geographically widespread spatial data used in the aforementioned GIS classifications can be further exploited to enable both predictive geographic infilling across space and estimation of specific covariate effects. Such approaches, however, rely on good spatial coverage of both the observation data and the predictor variables used to build the models. As many different forms of spatial environmental data (such as rasters) are becoming more accessible, and GIS tools become more ubiquitous, the development of methods which make best use of these data for environmental research is timely and of increasing importance in dealing with environmental change scenarios and providing appropriate advice to policy makers and environmental stakeholders.

The use of similar statistical regression modeling techniques, such as standard GLMs (McCullagh and Nelder, 1989), GAMs (Hastie and Tibshirani, 1990) and MARS (Friedman, 1991), has been common in both epidemiology and in species distribution / ecological niche modeling for some time. In the epidemiology literature such approaches are commonly used to map disease risk, incidence and spread (Vieira et al., 2005, Nguyen et al., 2012, French and Wand, 2004). In the ecology literature attention has been more focused on predictive modeling and understanding environmental effects rather than purely spatial analysis (e.g. Kriging or GIS classification). The mapping approach presented here demonstrates the use of a species distribution modeling regression approach with the inclusion of a spatial correlation structure (as we are ultimately interested in the spatial distribution). Although sometimes included when modeling and mapping individual species' distributions, this approach has rarely been applied specifically to the concept of mapping natural capital and indicators thereof.

In this paper, we present the application of a spatial statistical regression model using two national-scale data sets to explore the benefits of this approach against the use of simple environmental stratification. We reflect on how such approaches could be used to gain information on the distribution and extent of natural capital, and multiple environmental indicators across Great Britain.

Materials and Methods

National scale environmental data

The mapping of environmental indicators, either by GIS classification or statistical modelling, requires high quality observation data and covariate data with good spatial coverage (no obviously sparse areas) over the region of interest, preferably at high resolution with sufficient sample size. The Centre for Ecology and Hydrology (CEH) and the British Geological Survey (BGS) provide spatial information across Great Britain at 25 m and 50 m resolution on land-cover and parent material, respectively. Having national coverage of two key land-surface influences is important in determining the potential location of natural capital. Hence the two covariates can provide a solid basis for modelling and mapping natural capital and ecological responses to changes in land cover and parent material at the national scale. In the future, other covariates could be incorporated into the methodology, but for simplicity and as an example only two have been used in this paper.

The Land Cover Map 2007 (LCM2007) provides information about physical materials on the Earth's surface over the UK (Morton et al. 2011). Such physical materials may be manmade urbanised areas consisting of roads or buildings, or natural materials such as vegetation, exposed rock on inland water. The LCM2007, derived from satellite imagery, was produced as part of the Countryside Survey of the UK as a snapshot audit (Morton et al. 2011). Ground truthing and knowledge-based enhancements are also used to derive the physical coverage from the satellite images that make up the dataset, which is a continuous parcel-based (polygon) dataset accompanied by a suite of derived raster products with 25 m and 1 km resolution.

The Parent Material Model (PMM) is a spatial database representing below ground material from which the topsoil develops (Lawley, 2008). The PMM enables the distribution of physiochemical properties of the weathered and un-weathered parent materials to be mapped. It details over 30 rock and sediment characteristics adding simplified classifications of lithological properties. The attribute content includes a range of texture information, colour, structure, mineralogy, lithology, carbonate content and information about how the parent rock was formed (genetic origin) (British Geological Survey, 2013).

Natural Capital Data

As an example assessment of the possible benefit gained by adopting a geostatistical modelling approach over classification methods, we consider data on soil microbial community structure obtained from Countryside Survey (CS) 2007 (Norton et al. 2012). This dataset represents information on bacterial biodiversity at a nationwide extent. Soil bacterial biodiversity can be considered a good indicator unifying various parameters pertaining to natural capital, in that it is a biodiversity measure responsive to both natural fixed environmental factors such as geology and also changes in climate and land use (Griffiths et al., 2011). In a previous study analyzing these data, Griffiths et al., 2011 used a molecular approach (Terminal Restriction Fragment Length Polymorphism) to characterise the bacterial communities in soils from over 1000 cores sampled across Great Britain within the Countryside Survey sampling framework, which consisted of up to five randomly sampled soils taken from over 200 1-km^2 locations across GB. In their study, non-metric multidimensional scaling (NMDS) was used on the Bray-Curtis similarities of the community profiling results to define community composition in two dimensions. The first axis scores resulting from their ordination form the microbial community data used in the remainder of the work presented here.

The data were assessed by Griffiths et al., 2011 in relation to other environmental variables collected as part of the survey, including abiotic aspects of the environment as well as soil physical and chemical parameters. Those authors found that bacterial communities at this landscape scale were structured in similar manner to plants, and were highly correlated with a general gradient of soil parameters from acidic-organic soils to neutral soils of lower organic matter. This gradient was apparent in the first axis NMDS site scores, which generally increased with increased soil pH, and declining organic matter. These soil features are generally determined by the underlying geology and climate as well as associated human land usage. Therefore soil pH and plant biodiversity ordination scores were found to be amongst the best variables correlating with measures of bacterial biodiversity, but the aggregate vegetation classification (AVC) was also a strongly predictive factor.

To upscale the data from the discreet sampled locations and produce a GB scale map, Griffiths et al (2011) used the interpolation technique inverse distance weighting (Figure 1). Such a map is successful in illustrating the broad differences in communities between, for example, England and Scotland, but is unlikely to hold predictive power at smaller spatial scales. Here, we suggest that since vegetation cover and pH are strong predictors, and that the observed dataset has good spatial coverage due to the stratified sampling design of CS, we can use a more informative model-based approach to make more predictive spatial extrapolations. In particular we seek to test whether a more predictive spatial mapping can be obtained by using the LCM and PPM national coverage maps, compared to making naive use of an existing classification.

Statistical analysis

Given data on a numerical indicator of natural capital with suitable spatial coverage over the region in question, statistical models can be used to model the relationship between the indicator and other environmental covariate data. The model framework adopted needs to be flexible enough to cover the potentially complex structure of the observational data, whilst at the same time taking care to avoid false assumptions of independence, normality and linearity. An example of such a framework is the Generalised Linear Geostatistical Model (GLGM) of Diggle et al. (1998). This framework can easily be extended to a more generic setting where the linearity assumption is relaxed to form a Generalised Additive Geostatistical Model (GAGM) following on from the Generalised Additive Model framework (Hastie and Tibshirani 1990), which is already commonly adopted in species distribution modeling. The underlying model framework of a GAGM consists of three parts: 1) a linear combination of potentially smoothly varying covariate functions; 2) a spatial random field, which we will define as a Stationary Gaussian Process (SGP); and 3) random effects representing underlying, potentially non-spatial, error structure.

Having modelled the relationship between stock estimates of particular indicators reflecting national capital (such as: soil carbon; water quality; plant species occurrence; and in this instance soil microbial community structure) and the environmental covariates, one can, within the bounds of the training data, interpolate across unsampled geographic regions using information on the covariates available over finer spatial scales. For prediction of this sort it is essential that the observed data demonstrate both good spatial coverage and good covariate coverage such that predictions are not made beyond the range of this training data set—i.e. all geographic areas where we wish to make predictions are represented and the full range of covariate values are represented in the data set that the models were built on. In species distribution modeling, this is often referred to as the difference between analog and non-analog conditions (see for example Williams and Jackson, 2007; Veloz et al., 2012; Algar et al., 2009), where non-analog conditions are those unlike any previously observed in the study. Providing that the geographic and covariate space over which predictions are sought has a suitable analog in the observed data, substituting the wide coverage covariate data into the estimated model achieves predictions over the same spatial extent for the same snapshot in time as the observed response data. The geostatistical model-based approach of Diggle et al. (1998) has the clear advantage over simple kriging and GIS classification that both spatial correlation structure and covariate effects are taken into account. Furthermore, the model-based approach allows for simple extraction of the estimated error structure, and hence we can quantify the uncertainty in the predictions. Further details on the model framework including mathematical specification are provided in Supplementary Material Appendix 1.

In following this modeling procedure, we first carried out a GIS 'points in polygon' procedure to concatenate the CS data on microbial communities with corresponding data on land cover and calcium carbonate content. The final dataset consisted of 1010 observations. The raw data on soil microbial community ordination scores were modeled against broad habitat and calcium carbonate content using a generalised additive mixed-model (Lin and Zhang, 1999) approach. This follows the same generic approach as the GAGM without the inclusion of a spatial random field, which was deemed redundant upon examination of model residuals using Moran's *I*. The random components in the mixed model were needed to account for the apparent non-independence between any two soil cores taken from the same 1km square. These were more likely to be similar than two cores taken from two different squares. Alongside the random effects and fixed effects of habitat and calcium carbonate, an additional spatial surface was included to account for residual large scale spatial variation.

The model was fitted, including the smoothly varying spatial surface using tensor product smooth interactions, via the gamm function in the ‘mgcv’ library (Wood, 2011) in the R statistical environment (R Development Core Team, 2008). Estimates of the model parameters were obtained using restricted maximum. Full details of model specification and testing are provided in Supplementary Material Appendix 2, which also provides details on model fitting when the spatial random field is needed in the model formula.

For purposes of comparison, we then used the ITE land classification (Bunce et al., 1996) to produce a classification-based assessment. This was obtained by simply taking the mean microbial ordination axis score per land class. As the same land classification is used to classify the CS samples, sufficient sample size was guaranteed in each classification segment. What we are hence comparing is a model-based map versus the naive use of an existing geographic classification. Existing classification maps are often used in this way as it is not always feasible to develop a new classification for each purpose.

Examination of the mean square error of the predictions against the observed data provides a formal comparison of the goodness of fit of the model-based approach versus the classification-based approach. Mean square errors are obviously produced at an observation level, but here we wanted to map them to assess any spatial characteristics and areas where the model was and was not performing well. To do this the average mean square error in each habitat*calcium carbonate category was calculated (or land class category) and this value mapped according to where that category is present over GB.

Results

In the model-based approach, parameter estimates and associated P values of the fixed effects show a high degree of dissimilarity amongst the factor levels of each of the category values (Table 1). High levels of calcium carbonate content are correlated with high values of the microbial community metric. This is consistent with the findings in Griffiths et al. (2011) who showed a positive relationship with the community metric and pH. Likewise, the acidic habitats, such as dwarf shrub heath, coniferous woodland and acid grassland, show low values for the community score, again consistent with findings of those authors.

After estimating all unknown parameters in the relationship between microbial community structure on one hand and land cover and calcium carbonate content on the other, and checking these parameters against expert knowledge gained from previously published results, predictions were obtained over Great Britain by substituting the full LCM and PMM data into the equation from the fitted model together with the spatial coordinates (Figure 2C). Similar models and maps were produced for the two sub-models which contain a single predictor variable each: land cover OR calcium carbonate content (Figures 2A-B). This separation enables a visual inspection of effects of each specific covariate and is a clear advantage over the classification-based approaches where it is fully unknown what is driving the spatial pattern and how. Although informative with regards to specific covariates, the model is a correlative assessment and any robust inference on drivers of change is confounded by the possible correlation between covariates included the model and missing ones. Care is therefore needed when interpreting the estimated relationships between the response and individual model terms.

As an interpretation of the maps presented in Figure 2, it appears that the land cover data enable separation of the response between the upland and lowland dominated habitats (Figure 2A), a feature clearly visible in the Kriging-based map (Figure 1), whereas the calcium carbonate data allow separation of the lowland habitats into the alkaline and acidic soils

(Figure 2B). The maps produced also echo the findings in Griffiths et al., 2011 that both factors are required to adequately describe the spatial variation exhibited in microbial community structure (Figure 2C).

Comparison with classification-based approach

The classification-based map, derived using the ITE land classification (Bunce et al., 1996), uses colours on the same gradient scale as the model-based results to indicate the estimated mean within each class (Figure 2D). Comparing the full model-based map (Figure 2C) to the map drawn using classification means (Figure 2D), shows that although the two maps look broadly similar, it is unknown what key components make up the soil microbial community structure and what drives the spatial segregation in the classification-based map.

The mean square errors from each of the mapping approaches are mapped with the darker colours representing a lower mean square error and hence better goodness of fit (Figure 3). It is clear that the modelled approach of using both land cover and parent material provides the best fit to the data. It also shows how the model-based approach is more informative, by examining the relative contribution of each variable as layers are included or discounted in the model. Integrating the mean square error over the whole area provides a simple single statistic assessment and shows that the model-based map using land cover and geology provides the best fit (lowest total mean square error of 10482.60 versus 22929.54 for the classification-based map). The classification-based map, however, still provides some information, indicating potential areas where it provides a better fit than the model-based approach. An example here would be around The Fens in East Anglia (highlighted by the red box). Thus it is clear the model-based approach may be missing an important driving variable (or any correlate of that missing driver) that represents the differing microbial community structure found in this area.

Discussion

Understanding spatial trends in natural capital indicators and their relationships with environmental conditions is vital in supporting evidence-based policy. The example presented demonstrated a procedure to facilitate this by modeling and subsequently mapping one particular indicator of natural capital known to have a significant impact on terrestrial ecosystem functioning. Though it is tempting to use these types of models to draw inference on drivers of change and the causal pathway behind the current state of natural capital, they can only identify potential environmental drivers and the variables that show a clear relationship to the response. This is because the models themselves represent a correlative assessment to establish relationships present in the observation data. To understand the role of mechanistic drivers, an assessment involving experiments and specifically designed long-term studies is necessary (Holland, 1986). However, if the sole purpose of the analysis is prediction, as spatial mapping is, rather than understanding drivers of change, then any confounding correlation between included covariates and missing covariates is not critical (Araújo and Guisan, 2006).The example used only two covariate datasets, however, it would be trivial to add further environmental variables such as climate or topographical features. This would increase the flexibility of the model-based approach and is likely to reduce the mean square error further across the geographic range.

Previous work in this area has often focused on the use of classification-based maps to provide a framework onto which one can express the value of natural capital. The results showed that the model-based map outperformed the classification approach. In our particular example this was perhaps not surprising - Griffiths et. al. 2011 had already demonstrated land cover was a key factor in microbial community response, and land cover is omitted in the classification of Bunce et al (1996). Classification maps are often developed without the inclusion of variables that may be subject to change over time. This is to ensure that the geographic classification remains robust. Hence the exclusion of land cover occurs in many classifications, as it can be highly temporally variable. This example further highlights the issues surrounding naive use of existing classifications and why, given appropriate data, a model-based approach ought to be favoured.

The model-based approach to mapping natural capital presented, whilst extremely powerful and informative, relies heavily upon data with good spatial coverage, both in terms of the response one wishes to model and the variables with which to make prediction across a wider range of unsampled locations. It is therefore clear that coordinated, large scale, nationwide monitoring schemes such as the Countryside Survey (Norton et al., 2013), which play a pivotal role in providing source data on natural capital assets, should be maintained and exposed to inform policy decisions.

With increasing pressure on our natural assets from increasing human requirements and environmental change, there is an urgent need to provide better information for policy development and decision support. If we are to fully understand and value natural assets and ensure that they feed into decision-making, then it is important that we understand their distribution and trends in national extent and condition. Initiatives such as the Valuing Nature Network (VNN) and Natural Capital Committee (NCC) in the UK are government funded initiatives with the remit of ensuring that the national contribution of natural assets to a range of societal and economic benefits is well understood and helpfully informs decision making. This is done whilst balancing competing pressures and assessing the impact of different policy scenarios. Natural capital initiatives like the VNN and NCC also often seek to understand trade-offs and co-benefits across multiple environmental responses to help in conservation management, planning and resource distribution. We therefore anticipate that the powerful, information rich, model-based approach to understanding and mapping natural capital will increase in use over the coming years as we seek to value our natural assets and predict landscape scale responses to change in environmental or policy drivers.

Acknowledgements

The authors gratefully acknowledge the support of Simon Smart, Russell Lawley and Matthew Harrison in the review and editing of this paper. We also express significant thanks in particular to the associate editor and also to three anonymous referees for their helpful comments and suggested edits that resulted in a significantly improved revision of this paper. This paper has been published as a result of research conducted by the CEH and BGS through their NERC directed research programmes. In addition, RIG was supported by the European Commission under the EcoFINDERS project (FP7-264465). The authors publish this article with the permission of the Executive Director, British Geological Survey (BGS), Natural Environment Research Council (NERC).

References

Algar, A. C., Kharouba, H. M., Young, E. R., & Kerr, J. T. (2009). Predicting the future of species diversity: macroecological theory, climate change, and direct tests of alternative forecasting methods. - Ecography, 32(1): 22-33.

Araújo, M. B., and Guisan, A. 2006. Five (or so) challenges for species distribution modelling. – Journal of Biogeography, 33(10): 1677-1688.

Bunce, R. G. H. et al. 1996. Land Classification for Strategic Ecological Survey. – Journal of Environmental management 47: 37-60.

British Geological Survey. 2013. Soil parent Material Model. Available online at: http://www.bgs.ac.uk/products/onshore/soilPMM.html. [Cited 07 February 2012].

Carpenter, S.R. et al. 2009. Science for managing ecosystem services: Beyond the Millennium Ecosystem Assessment. – Proceedings of the National Academy of Sciences of the United States of America. 106(5): 1305–1312.

Costanza, R., and Daly, H. E. 1992. Natural Capital and Sustainable Development. – Conservation Biology. 6 (1): 37-46.

Costanza, R., Wilson, M., Troy, A., Voinov, A., Liu, S., and D'Agostino, J. 2006. The value of New Jersey's ecosystem services and natural capital. – Gund Institute for Ecological Economics, University of Vermont and New Jersey Department of Environmental Protection, Trenton, New Jersey, 13.

Diggle, P. J. et al. 1998. Model-based geostatistics (with discussion). – Journal of the Royal Statistical Society, Series C. 47(3): 299-351.

Friedman, J. H. 1991. Multivariate adaptive regression splines. – The annals of statistics, 1-67.

Griffiths, R. I. et al. 2011. The bacterial biogeography of British soils. – Environmental Microbiology. 13 (6): 1642-1654.

Hastie, T. J. and Tibshirani, R. J. 1990. Generalized Additive Models. – Chapman and Hall, New York

Holland, P. W. 1986. Statistics and causal inference. – Journal of the American statistical Association, 81(396): 945-960.

Lawley, R. 2008. The soil-parent material database: A user guide. – British Geological Survey Internal Report .OR/08/034. 45pp.

Lin, X. and Zhang, D. 1999. Inference in generalized additive mixed models by using smoothing splines. – Journal of the Royal Statistical Society, Series B. 61: 381-400

McCullagh, P. and Nelder, J.A. 1989. Generalized Linear Models, 2nd edition. London: Chapman and Hall/CRC Press

Morton, D. et al. 2011. Land Cover Map. 2007. Countryside Survey: Final Report for LCM2007 – the new UK Land Cover Map. – Centre for Ecology and Hydrology. CS technical Report No 11/07.

Millennium Ecosystem Assessment. 2005. Ecosystems and Human Well-being: Synthesis. – Island Press. Washington DC.

Naidoo, R and Ricketts T. 2006. Mapping the economic costs and benefits of conservation . PLoS Biology. 4 (11): doi:10.1371/journal.pbio.0040360.

National Research Council. 2005. Valuing ecosystem services: Toward better environmental decision-making. National Academy Press, Washington, DC.

Nengwang, C. et al. 2009. A GIS-based approach for mapping direct use value of ecosystem services at a county scale: Management implications. – Ecological Economics. 68(11): 2768-2776.

Norgaard, R. B. 2009. Ecosystem services: From eye-opening metaphor to complexity blinder. – Ecological Economics. 69 (6): 1219-1227.

Norton, L. R., Inwood, H., Crowe, A., and Baker, A. 2012. Trialling a method to quantify the ‘cultural services’ of the English landscape using Countryside Survey data. – Land Use Policy. 29(2): 449-455.

Norton, L. R. et al. 2013. Measuring stock and change in the GB countryside for policy - Key findings and developments from the Countryside Survey 2007 field survey. – Journal of Environmental Management. 113: 117-127.

R Development Core Team. 2008. R: A language and environment for statistical computing. – R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Raymond, C. M., Bryan, B. A., MacDonald, D. H., Cast, A., Strathearn, S., Grandgirard, A., and Kalivas, T. 2009. Mapping community values for natural capital and ecosystem services. – Ecological economics. 68(5): 1301-1315.

Sayre, R. et al. 2009. A new map of standardized terrestrial ecosystems of the conterminous United States. – U.S. Geological Survey Professional Paper 1768, 17 p. (Also available online – pubs.usgs.gov/pp/1768.)

Swetnam, R. D.et al. 2011. Mapping socio-economic scenarios of land cover change: A GIS method to enable ecosystem service modelling. – Journal of Environmental Management. 92: 563-574.

Troy, A. and Wilson, M. A. 2006. Mapping ecosystem services: practical challenges and opportunities in linking GIS and value transfer. – Ecological economics. 60(2): 435-449.

Turner, R. K., et al. 2010. An Introduction to socio-economic assessment with a marine strategy framework CSERGE, UEA – Published by the Department for Environment, Food and Rural Affairs.

The UK National Ecosystem Assessment. 2011. The UK National Ecosystem Assessment: Synthesis of the Key findings. – UNEP-WCMC, Cambridge.

Veloz, S. D., Williams, J. W., Blois, J. L., He, F., Otto-Bliesner, B., & Liu, Z. 2012. No-analog climates and shifting realized niches during the late quaternary: implications for 21st-century predictions by species distribution models. – Global Change Biology, 18(5): 1698-1713.

Williams, J. W. and Jackson, S. T. 2007. Novel Climates, No-Analog Plant Communities, and Ecological Surprises: Past and Future. – Frontiers in Ecology and the Environment 5: 475-482

Wood, S.N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. – Journal of the Royal Statistical Society (B). 73(1): 3-36

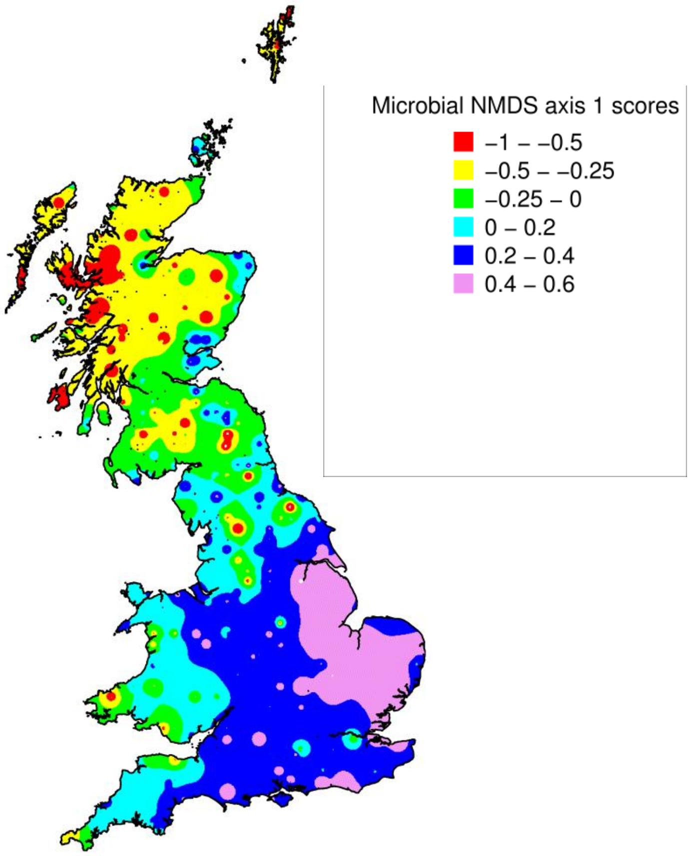

Figure 1: Map of soil microbial community structure (NMDS first axis scores) based on kriging of data obtained from the Countryside Survey – a stratified random sample of 591 1km survey squares located across the whole of Great Britain.

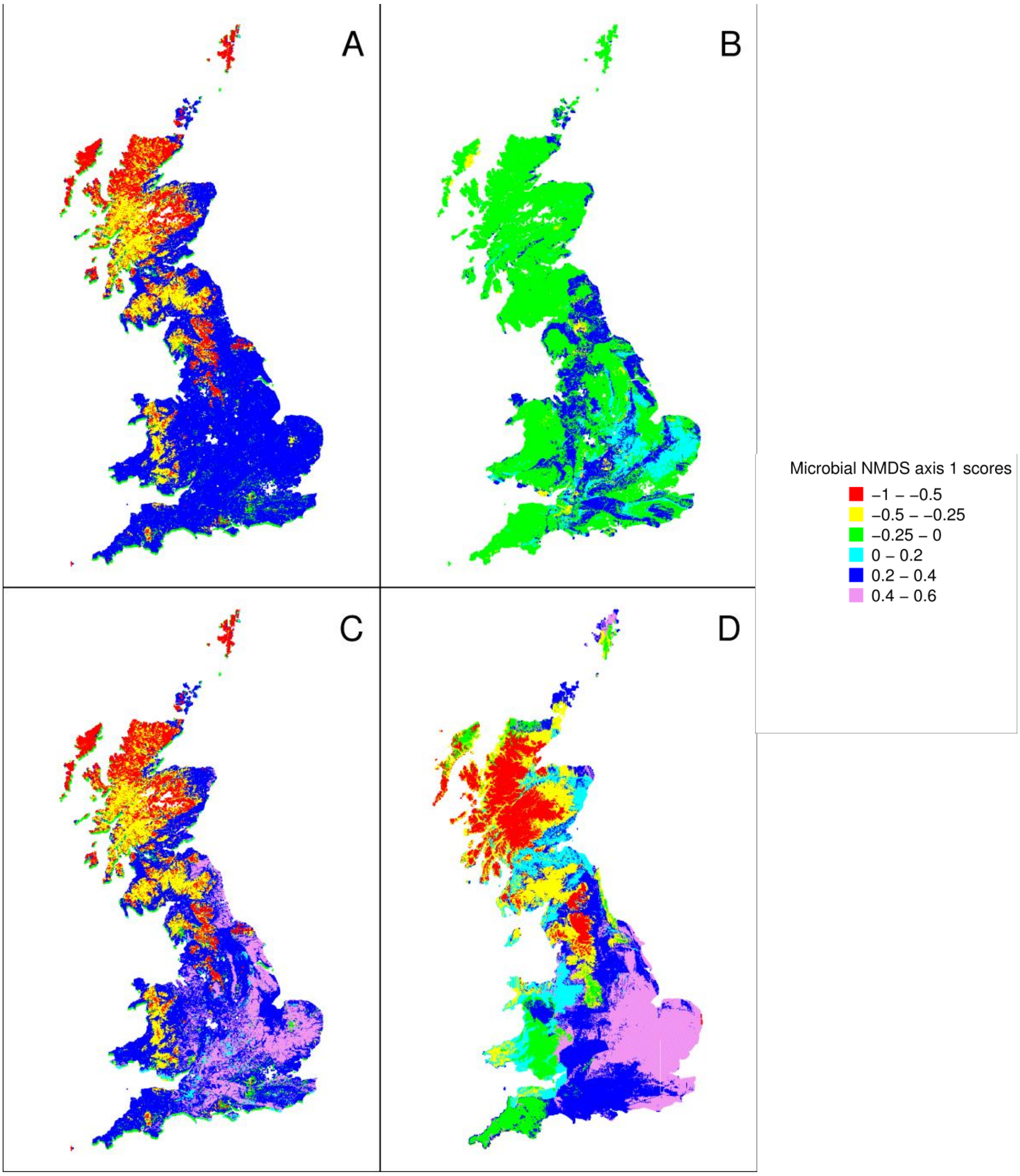

Figure 2: Maps of predictions in soil microbial community structure over Great Britain at 1km resolution, showing comparisons among covariates of the model-based approach, and contrasting results of the model-based and classification analyses. A - C using model-based approaches with covariates: A) land cover only; B) calcium carbonate content only; and C) land cover and calcium carbonate content combined. D estimating mean levels in each environmental stratum defined by the ITE land classification of GB, then displaying on map using the spatial outline of each stratum.

Figure 3: Goodness of fit of the spatial statistical model used to derive the relationship between soil microbial community structure and environmental variables (land use and calcium carbonate content of the soil parent material). Map shows mean square error in each of the land use*calcium carbonate classes. Darker shades indicate areas with low error.

Table 1: Estimated parameters and associated standard errors and p values resulting from the spatial statistical model estimated defining the relationship between soil microbial community scores and environmental variables (land use type and calcium carbonate class).

Parameter	Estimate	Standard Error	P Value
Intercept (CACO3 VARIABLE(LOW) * Bog)	-0.45	0.04	< 0.001
CACO3 HIGH	0.16	0.05	< 0.001
CACO3 LOW	-0.11	0.04	0.004
CACO3 MODERATE	0.04	0.14	0.786
CACO3 NONE	-0.10	0.03	< 0.001
CACO3 UNKNOWN	-0.25	0.14	0.078
CACO3 VARIABLE	-0.01	0.04	0.848
CACO3 VARIABLE(HIGH)	-0.31	0.10	0.002
Broadleaved, Mixed and Yew Woodland	0.46	0.04	< 0.001
Coniferous Woodland	0.10	0.04	0.011
Arable and Horticultural	0.87	0.03	< 0.001
Improved Grassland	0.77	0.03	< 0.001
Neutral Grassland	0.68	0.03	< 0.001
Calcareous Grassland	0.60	0.12	< 0.001
Acid Grassland	0.12	0.03	< 0.001
Bracken	0.24	0.07	0.001
Dwarf Shrub Heath	-0.01	0.04	0.818
Fen, Marsh, Swamp	0.46	0.06	< 0.001

Supplementary Material

Appendix 1

A model framework suitable for spatial modelling and mapping is the Generalised Linear Geostatistical Model (GLGM) of Diggle et al. (1998). This framework can easily be extended to a more generic setting where the linearity assumption is relaxed to form a Generalised Additive Geostatistical Model (GAGM) following on from the Generalised Additive Model framework (Hastie and Tibshirani 1990). Both model frameworks allow for the key relationships to be estimated between the response of interest and the environmental covariates, whilst at the same time controlling for additional spatial effects. This is because observations close to one another are more likely to be similar than observations far away, even after accounting for the environmental covariates in the model.

Spatial autocorrelation can be accounted for by including a purely spatial term in the model, often a spatial random field, which captures any residual spatial variation in the data. This ensures that parameter estimates and their associated standard errors are unaffected by any residual spatial dependence. It also has the advantage that one can use the estimated spatial correlation structure when making predictions, thus maximising the use of information, in an approach similar to simple kriging. The underlying model framework of the GAGM considered is presented below, where the geostatistical model consists of three parts: 1) a linear combination of potentially smoothly varying covariate functions; 2) a spatial random field, which we will define as a Stationary Gaussian Process (SGP); and 3) random effects representing underlying, potentially non-spatial, error structure. Mathematically the model framework is represented as

$$E[Y_i] = g\{\eta_i\}$$

$$\eta_i = \alpha + \sum_{j=1}^{k} f_j(x_{ji}) + \boldsymbol{S}(u_i) + \mathbf{Z}_i\mathbf{b} \qquad (1)$$

where Y is the response variable, f_j are smooth functions (generally cubic regression splines) of environmental covariates x_j, g is the link function (as with standard GLMs), α is the intercept term, **Z** represents different grouping levels, b ~ N(0 , σ) represents the differing variation assigned to each of the groups in **Z** and ***S*** is a Stationary Gaussian Process at location u_i with zero mean and covariance structure given by $\mathrm{Cov}(u, u') = \sigma^2\rho(||u - u'||)$.

Appendix 2

As with all statistical modelling approaches it is more appropriate to start with a model consisting of a fixed effects formula dictated by scientific understanding and a simple error structure. Then, upon testing residuals and model assumptions, adapt the error structure as necessary. In this example we hence started with a simple GAM with land cover and calcium carbonate data as predictor variables together with a purely spatial interaction term of latitude and longitude to account for large scale spatial effects. Fitting a spatial trend surface is crucial to ensure adequate attribution of the response to the model covariates (Legendre and Fortin, 1989).

Upon examination of the residuals, it was clear that within square variance was not the same as the between square variation; hence the assumption of independence in the residuals was flawed. We therefore re-fitted the model with a random intercept effect to account for which CS 1km square the soil data were obtained from. This allowed for small scale random adjustments in the model. The residuals from the re-fitted model did not appear to imply any heteroscadacity or any obvious key missing hierarchy in the error structure.

The residuals were then analysed for any small scale spatial autocorrelation. This was done using Moran's I, which showed no signs of small scale spatial autocorrelation apparent in the residuals. As this spatial autocorrelation was assessed on the residuals there was no need to include any disconnection when calculating Moran's I as any differences should have been accounted for in the main effects. Previous studies (eg Franklin and Mills, 2003) have shown spatial autocorrelation of soil microbial community data is evident at distances of up to 7 metres. CS squares are separated by a minimum of 15 km and within square observations are separated by a minimum of 80 metres with an average separation distance of 558 metres. Given this, and the results of Franklin and Mills, the redundancy of fine scale spatial autocorrelation in the model is perhaps not surprising.

We therefore modelled the raw data on soil microbial communities against broad habitat and calcium carbonate content using a generalised additive mixed-model based approach. This follows the same generic approach as the GAGM without the inclusion of a spatial random field. Generalised additive mixed models (Lin and Zhang, 1999) extend the framework of the standard GAM by allowing both fixed and random affects to be present in the model. The random components can account for unobserved affects that could influence the outcome of the response variable and therefore ensure that estimated standard errors are accurate and any inference is reliable. Extending the general GAM equation to include random effects gives us a model of the following form:

$$g(E[Y_i|\boldsymbol{x},\mathbf{b}]) \;=\; \alpha + \sum_{j=1}^{k} f_j(x_{ji}) + \mathbf{Z}_i\mathbf{b}$$

where y is the response variable, f_j are smooth functions (generally cubic regression splines), g is the link function (as with standard GLMs), α is the intercept term, Z represents different grouping levels and b ~ N(0 , σ) represents the differing variation assigned to each of the groups in Z.

The random components are used here to allow us to account for the fact that any two soil cores taken from the same 1km square are more likely to be similar than two cores taken from two different squares. The non-linear smooth form allows fitting of an additional smoothly varying spatial surface to soak up any residual large scale spatial variation and hence captures the spatial structure present in the data that our covariates may not adequately explain. This is akin to including time as a covariate in time series modelling – the user is effectively de-trending the data. Even in the absence of small scale spatial autocorrelation, Legendre and Fortin (1989) emphasised the importance of including this term. Including the random effects, additional spatial surface and the habitat and calcium carbonate covariate effects, the fitted model is thus represented by

$$g\{\mathbf{E}[smc_{i,s}]\} = \alpha + \beta_{h_i} + \eta_{c_i} + f(Latitude_i, Longitude_i) + \omega_s + \sigma_i$$

where for each observation i in square s, smc is the soil microbial community score, β_h is the estimated value of habitat h associated with observation i, η_c represents the value for calcium carbonate category c, ω represents the error (normally distributed) associated specifically with square s and σ represents the residual model error also assumed to follow a normal distribution. The model was fitted, including the smoothly varying spatial surface using tensor product smooth interactions, using the gamm function in the 'mgcv' library (Wood, 2011) in the R statistical environment (R Development Core Team, 2008).

Had the re-fitted model failed the independence assumptions and the Moran's I test showed evidence for fine scale spatial autocorrelation, then the inclusion of the spatial random field term in the model would have been necessary. Practically, the Gaussian Random Field (GRF) is often estimated by making the assumption that it is adequately specified by a Markov Random Field (MRF) whereby each location only depends on its "neighbours" and is conditionally independent of all other locations. The neighbourhood structure of the MRF allows the spatial component of the model to be estimated by methods such as Conditional Autoregressive Models (CAR) or Simultaneous Autoregressive Models (SAR). Dormann et al (2007) provide an overview of methods for accounting for spatial autocorrelation including description of CAR and SAR models and how to fit them in practice with clearly referenced R packages.

It is worth noting that both CAR and SAR models can also be estimated in a Bayesian framework, where estimated parameters and standard errors are often more reliable than in likelihood approximation methods, though with an added computational cost. The advantage is the added flexibility that moving to the Bayesian paradigm brings. Specifically in this case the possible inclusion of smoothly varying penalised regression splines following the approach taken by Crainiceanu et. al. (2005). This provides the full ability to fit the model specified in Eqn (A1). This type of model can also be easily fitted using Integrated Nested Laplace Approximation (Rue et. al., 2009), where robust parameter estimates can be obtained quickly and efficiently. The R package R-Inla (www.r-inla.org) is a user friendly resource for fitting the model in Eqn (A1) using this approach.

References

Crainiceanu, C. M., Ruppert, D. And Wand, M. P. 2005.Bayesian Analysis for Penalized Spline Regression Using WinBUGS. – Journal of Statistical Software. 14

Diggle, P. J. et al. 1998. Model-based geostatistics (with discussion). – Journal of the Royal Statistical Society, Series C. 47(3): 299-351.

Dormann, C. F. et al. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. – Ecography. 30: 609-628.

Franklin, R. B. and Mills, A. L. 2003. Multi-scale variation in spatial heterogeneity for microbial community structure in an eastern Virginia agricultural field. – FEMS Microbiology Ecology, 44: 335–346.

Hastie, T. J. and Tibshirani, R. J. 1990. Generalized Additive Models. – Chapman and Hall, New York

Fortin, M. J., Drapeau, P., & Legendre, P. 1989. Spatial autocorrelation and sampling design in plant ecology. – Vegetatio, 83(1-2): 209-222.

Lin, X. and Zhang, D. 1999. Inference in generalized additive mixed models by using smoothing splines. – Journal of the Royal Statistical Society, Series B. 61: 381-400

R Development Core Team. 2008. R: A language and environment for statistical computing. – R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rue, H., Martino, S. & Chopin, N. 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. – Journal of the Royal Statistical Society, Series B. 71(2): 319-392.

Wood, S.N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. – Journal of the Royal Statistical Society (B). 73(1): 3-36