BTO Research Report No. 318

Producing statistically valid maps of species abundance from UK Breeding Bird Survey counts using Geostatistical Analyst in ArcGIS

Authors

S.E. Newson and D.G. Noble

A report by the British Trust for Ornithology

April 2003

British Trust for Ornithology, The Nunnery, Thetford, Norfolk, IP24 2PU
Registered Charity No. 216652

S.E. Newson and D.G. Noble

Producing statistically valid maps of species abundance from UK Breeding Bird Survey counts using Geostatistical Analyst in ArcGIS

BTO Research Report No. 318

A report by the British Trust for Ornithology

Published in March 2005 by the British Trust for Ornithology
The Nunnery, Thetford, Norfolk IP24 2PU, UK

ISBN 1-904870-25-2

CONTENTS

1. SUMMARY

We examine the potential of the Geostatistical Analyst extension of ArcGIS for interpolating statistically valid maps of species abundance from survey data. To explore this methodology, we use Breeding Bird Survey (BBS) data for 2000, covering 11 species ranging from widespread and abundant to rare and localised species.

The results demonstrate that it was possible to produce maps that matched well the expected distribution and abundance for the majority of species. However it was not possible to produce maps for Willow Tit and Nightingale, which are poorly monitored by the BBS because they occur at low densities and are highly localised in their distribution. Further to this, predictions of abundance for species that have specific habitat requirements and show a restricted range, such as Reed Warbler and Nuthatch based purely on location, are likely to be improved by narrowing the area over which predictions are made, and may benefit from co-kriging models which include habitat as a predictor variable. Alternatively presence/absence could be modeled using indicator kriging.

Examining the potential of this methodology for producing automated production of maps it was encouraging to find that models with default parameters chosen by the program compared well with predictions from manual diagnoses of the data and modelling. However, there is some reduction in the level of precision that will reduce the number of species for which abundance maps can be produced.

In addition to co-kriging and indicator kriging mentioned above, further work could use this methodology to model the temporal as well as spatial change in species abundance or distribution, providing a means of visually identifying geographic areas of significant population change, perhaps prior to further data analysis.

2. INTRODUCTION

The Breeding Bird Atlas of 1988-91 presented maps of species abundance for all abundant and widespread bird species in Britain and Ireland at that time (Gibbons *et al*. 1993). Abundance maps of this type are of huge importance, not only in highlighting the strongholds of particular species and through change maps allow areas of significant population change to be identified, but they allow information such as this to be made accessible to much wider audience than would normally be possible.

In the Atlas a deterministic interpolation method was used, which like all interpolation methods is based on the assumption that surveyed sites that are close to one another are more alike than those that are further apart. This was performed in the Atlas by weighting points closer to the prediction location greater than those further away (see Johnston *et al*. 2001 for a discussion of deterministic interpolation methods). However, over the last ten years since these maps were produced there have been considerable advances in the application of geostatistics to improve the estimation and precision of interpolated surfaces and the integration of advanced geostatistics within a GIS framework, most notably as implemented by the Geostatistical Analyst extension of ArcGIS (Johnston *et al*. 2001).

Geostatistical methods are based on statistical models that model autocorrelation (statistical relationship among measured points). Not only do these techniques have the capability of producing a prediction surface, but they can also provide some measure of the accuracy of the predictions. A number of geostatistical interpolation techniques have been produced, of which kriging is the most applicable to this project. Like the deterministic methods described above, kriging weights the surrounding measured values to derive a prediction for unsurveyed locations. However, the weights are not only based on the distance between measured sites and the prediction location, but also on the overall spatial arrangement in the weights, the spatial autocorrelation. For a full discussion of geostatistics and geostatistical methods see Chiles & Delfiner (1999).

In this project we examine the potential of recent software advances, in the particular the Geostatistical Analyst extension of ArcGIS (Johnston *et al*. 2001), to produce interpolated abundance maps from Breeding Bird Survey (BBS) data. In particular this report aims to identify the best approach for modeling BBS count data to produce predictions of spatial variation in abundance. This report also addresses the limitations imposed by making simple assumptions to allow for automated map production. Because we are interested in evaluation of the methods only here, we have concentrated on Britain. However, we explore the effect of introducing data for Ireland on the resulting maps and of producing separate maps for Northern Ireland using two species as examples, Wren and Meadow Pipit.

3. METHODS

3.1 Data preparation

To examine the questions above, we apply the methodology to eleven bird species recorded on BBS squares in 2000, ranging from abundant and widespread to rare and highly localised species (Table 1). In this year, a total of, 2160 1-km BBS squares were surveyed in Britain and Ireland, the geographical spread of which is shown in Figure 1.

From the raw BBS data, species-specific files were created in SAS (SAS 1996). Each file contained a list of all 1-km squares surveyed in 2000 and a count for each square, calculated as the sum over 200-m transect sections, distance categories (0-25-m, 25-100-m, 100-m or more and in flight) and two survey visits. In this study we exclude BBS squares surveyed on only one visit in this year. Squares where the species was not recorded were assigned a zero count. Full description of the survey design can be found elsewhere (Gregory *et al*. 1996, Gregory & Baillie 1998).

3.2 Modelling approach

Making accurate predictions

The Geostatistical Analyst extension of ArcGIS allows for a number of different types of geostatistical model with different assumptions and aims to be fitted to BBS data. However, because the BBS employs a stratified sampling design that results in unequal representation of coverage in different areas of Britain, we feel it necessary to control for this in the analyses, for which a kriging method known as simple kriging is used (see Johnston *et al*. 2001, pp131-166 for a comparison of different kriging methods). To control for variation in observer coverage, Geostatistical Analyst uses the method of declustering, which preferentially weights the count data, with counts in densely sampled areas receiving less weight and counts in sparsely sampled areas receiving greater weight (see Isaaks & Srivastava 1989 for a further discussion of this method). This effectively decides how much the data at each site contributes to the calculation of autocorrelation functions across the entire data set.

In Geostatistical Analyst there is a choice of two declustering methods that can be used i) cell declustering, which arranges rectangular cells over BBS squares in a grid and weight attached to each BBS square is inversely proportional to the number of BBS squares in its cell (Figure 2a) or b) polygonal declustering, which weights each BBS square in proportion to the areas that it represents (Figure 2b). In this study, we choose the first method in preference to the second, because with the second, it is likely to be difficult to define weights towards the coastline of Britain. Geostatistical Analyst chooses the optimal grid size, although a comparison with different gird sizes on the predictive error can be examined. It should be pointed out that although several geostatistical methods require that the data be normally distributed, prediction maps do not require this assumption to be met. BBS count data is unlikely to ever be normally distributed because there are a substantial proportion of zero counts. Ideally perhaps the best approach for BBS data would be use weighted polygons based on sampling regions, although this is not possible within the program.

Making automated predictions

One of the aims of this project is to examine the potential of this methodology and software for producing automated maps of predicted abundance. To obtain the most accurate predictions involves a number of stages to analyse the data and based on these findings make the best modeling decisions. However, because geostatistics allows us to calculate the level of predictive error associated with any predictions, we can make comparisons between these models and make a visual comparison of the resulting prediction maps.

3.3 Prediction maps

There are a number of stages involved in the production of an accurate prediction map. These include identifying and modelling global patterns in the data if they exist, understanding spatial autocorrelation and directional influence at a local scale and testing how well the chosen model makes predictions for unsurveyed locations. Haining (1990) provides a full discussion of the theory, while Johnston *et al*. (2001) explains how it can be applied using Geostatistical Analyst. The following is a brief summary of what is needed to understand the production of prediction maps, as applied to BBS data.

Identifying global patterns

If a global pattern (trend) exists in the count data, it may be represented by some mathematical formula. For example, a species with low abundance in the south increasing to high abundance in the north might be represented by a plane, whilst high abundance in the far south and north only, might be represented by a formula that creates a U shape (a second order-polynomial). However, in reality the formula is often too smooth to accurately depict the surface because no surface is a perfect plane or 'U' shape. If it is believed that the trend does not adequately portray the surface, it can be removed completely, leaving the short-range variation in the surface to be modeled (See Johnston *et al*. 2001 pp131-166). If the trend is believed to be valid and important for the prediction of abundance, the global trend is removed temporarily to allow for the short-range variation to be modeled, but added back before the final surface is created. As an example, the projected trend of Meadow Pipit in Britain is shown in Figure 3. This shows an increase in abundance from south to north and from east to west of Britain.

Understanding spatial autocorrelation and directional influences

To examine the spatial autocorrelation between counts on neighbouring BBS squares we use what is know as a semivariogram. The semivariogram is a function that relates semivariance (or dissimilarity) of count data on BBS squares to the distance that separates the squares (in fact the difference-squared of the values between pairs of locations at different distances). Its graphical representation can be used to show spatial correlation of count data on BBS with their neighbours. Geostatistical Analyst calculates the optimal parameters for a semivariogram model (a spherical model is chosen by default, where the best fit is in all directions) and determines the size of distance class (known as lag size) into which pairs of BBS squares are grouped and compared (known as binning). A number of other types of semivariogram model were examined, although these had very little effect on the resulting predictions and associated error. good lag size can help reveal spatial correlations and examines spatial autocorrelation between counts on neighbouring BBS squares at this scale. As illustrated in Figure 4 using the Wren as an example, Geostatistical Analyst produces a semivariogram graph and semivariogram surface. The colour scale on the surface represents a direct link between the empirical semivariogram values on the graph and those on the surface (lower values are blue and green and higher values orange and red. The x-axis on the semivariogram graph is the distance from the centre of the cell to the centre to the semivariogram surface. The semivariogram values represent dissimilarity. In the example here, the semivariogram starts low at small distance (things close together are more similar) and increases as distance increases (things get more dissimilar father apart). Notice also from the semivariogram surface that dissimilarity increases more rapidly in the southeast to northwest direction than in the southwest to northeast direction, so it appears that in this example there are directional components to the autocorrelation at a fine scale. The reasons for these directional influences are not known in this case, but they can be statistically quantified and will affect the accuracy of the surface that is created.

Searching neighbourhood

It is common practice to limit the data used for predictions of abundance to encompass a circle or ellipse around the point that predictions are being made. Additionally to avoid bias in a particular direction, the circle (or ellipse) is divided into a number of sectors. Assuming that surveyed BBS squares closer to where we are trying to predict a count are likely to be most similar, a weighting system is devised, where larger the weight, the most impact a particular location will have on the prediction at an unknown site. Geostatistical Analyst automatically chooses an optimum search neighbourhood based on the dataset, although this can be changed manually and the effect on the error rate determined using cross-validation discussed below.

Cross-validation

Cross-validation is used to give some idea of how well the model predicts at unknown sites. Using data from all BBS squares, cross-validation sequentially omits a square, predicts the count on that square using the rest of the data, and then compares the measured and predicted values. The calculated statistics serve as diagnostics that indicate whether the model is reasonable for map production.

An accurate model would have a mean error as close as possible to 0, the root-mean-square error and average standard error should be as small as possible (this is useful when comparing models), and the root-mean-square standardized error should be close to 1. Prediction error here is the difference between the prediction and the actual measured value.

4. RESULTS

4.1 Sample size restrictions

It was possible to produce maps of abundance for nine of the eleven species examined in this study (see Figure 5 a-i). The two species for which it was not possible to produce maps were Nightingale and Willow Tit occurring on 27 and 49 BBS squares respectively in 2000. However, whether a species can or cannot be mapped is not necessarily related to sample size, because it was possible to produce a reasonable map of abundance for the Ring-necked Parakeet occurring on only 12 BBS squares in 2000 (Figure 5d). The main difference with the Ring-necked Parakeet and these species is that all Ring-necked Parakeet records are from a very small geographic area and involve several birds at each site. Whilst in contrast, Willow Tit and Nightingale mainly involve records of single individuals over a larger geographic area.

For species that are highly restricted in their range, we can improve the predictions for the area in which the species occurs, by reducing the area over which we make predictions. This is illustrated with the Reed Warbler example in this study (Figure 5g). By excluding Scotland from the analyses, where the species is absent, we improve the prediction of abundance in areas in which this species occurs.

4.2 Reliability of abundance maps

Based on what we know of species abundance and abundance maps in the last Breeding Atlas (Gibbons *et al*. 1983), visual examination of maps produced here for the nine species seem to match very well what we would expect for these species. For example, abundance of Feral Pigeon is higher in cities (Figure 5a), whilst Wood Pigeon abundance is lowest in upland area of Wales, Scotland and the Peak District (Figure 5b). We see the highest abundance of Collared dove in suburban and rural areas around central London, in Norfolk where the species was first recorded and spread out from and around Birmingham and Manchester (Figure 5c), whilst Ring-necked Parakeets are restricted to south London districts (Figure 5d). Meadow Pipit abundance is highest in upland areas of Scotland and Wales (Figure 5e), Wren abundance is highest in England (Figure 5e), whilst Nuthatch are found at their highest abundance in Wales and the south and southwest of England (Figure 5f). House Sparrow shows the highest abundance in area of human habitation, but is not restricted to urban habitat as is seen in the Feral Pigeon (Figure 5g).

With each prediction map, it is possible to produce a map of standard errors across the predicted range (shown in Figure 6 a-i). Because these models predict abundance by location only, there are likely to be sites unsuitable for the species, close to suitable sites. This explains the higher deviation from the predicted in these standard error maps in those areas where the species is most abundant.

Predicted abundance maps for the Wren and Meadow Pipit including data for Ireland are shown in Figures 7 a & b and separately for Ireland in Figures 8 a and b. Although it is not sure the extent to which patterns of spatial autocorrelation in data from the UK mainland might contribute to the Northern Ireland results, visually the influence appears to unimportant.

4.3 Automated maps of abundance

A visual comparison of maps produced using default parameters chosen by Geostatistical Analyst and the best predictions by manual diagnoses and modeling of the data, suggest that default models performed well for the majority of species in this study. In addition to this, comparison of the average standard error associated with these two types of model in Table 2, suggests that the best predictions may result in little improvement over the predictions of the default model, although this varies between species.

However, because there is some reduction in precision using default parameters, it was now not possible to produce maps for the Ring-necked Parakeet and Nuthatch (Figures 9d & h) and the default map for Reed Warbler was visually different from the best prediction (Figure 9 g). Therefore, using default parameters may restrict the number of species for which maps can be produced.

5. DISCUSSION

This study highlights the potential of geostatistics and the Geostatistical Analyst extension of ArcGIS used here for producing statistically valid maps of species abundance for widespread and abundant bird species in Britain using BBS data. The maps provide a good match to the expected distribution of these species in Britain and to the predicted abundance maps presented in the last Breeding Atlas (Gibbons *et al*. 1993). However it was not possible to produce maps of abundance for the Nightingale and Willow Tit that occur at low densities and are highly localised in their distribution and are in fact poorly monitored by the BBS.

Because models in this study predict abundance using information on location only, and do not take into account the habitat or other requirements of the species. For a species, such as the Nightingale, which is localised and highly habitat-specific, it may be possible to map abundance for this species, if we use an independent dataset, containing information on the distribution of say coppice woodland as a predictor of abundance in what known as a co-kriging model. In reality, most species in Britain show some form of habitat or altitudinal preference, so this approach may be worth exploring further to examine the extent to which our predictions can be improved for a larger number or all species routinely monitored by the BBS. Co-kriging would involve little extra work within the program itself, although it would be first necessary to decide on the important predictor variables to include in the model. An alternative approach where there is limited data could be to model presence/absence using another geostatistical method known as indicator kriging. For a discussion of the theory relating to co-kriging and indicator kriging see Johnston *et al*. (2001). Additionally for species with a restricted range in Britain, such as the Reed Warbler example here, predictions may be improved by restricting predictions to exclude areas where the species is absent.

Before, maps of this type are to be produced, it would need to be decided whether abundance should be modeled separately for Britain and Northern Ireland. The combined maps in this study for the Wren and Meadow Pipit suggest that patterns of spatial autocorrelation from UK mainland are likely to have little influence on the resulting combined maps, so this may not be a problem. However, it maps are to be produced for Northern Ireland, it would make sense to combine this with data for Southern Ireland now available through the Countryside Bird Survey coordinated by Bird Watch Ireland, which uses the same methodology as the BBS.

In terms of the potential for automating this methodology, it was encouraging to find that models producing using default parameters chosen by Geostatistical Analyst produced a pretty good comparison with the best prediction made through manual diagnoses of the data and modeling. However, because there is some loss of precision when using an automated approach, it is likely to reduce the number of species for which it is possible to produce abundance maps. To fully automate the process with ArcGIS, time would need to be allocated to the development of a macro to perform this function in visual basic.

Other areas of study that would be interesting to pursue using this methodology include modeling the temporal as well as spatial change in species abundance or distribution. This would allow one to quickly identify and visualise areas in Britain in which there has been significant population change, possibly prior to further data analyses. There is also potential for a more rigorous examination of the error associated with the predictive surface, perhaps using a validation method to use part of the data to predict abundance for the remaining data in the model, although this is likely to be possible only for species most widely recorded by the BBS.

It is difficult to extrapolate from this study the number of species monitored by the BBS for which maps of abundance could be routinely produced. However, if one were to assume that maps could be produced for all species recorded on 50 or more BBS squares in any one year, it should be possible to produce maps for at least 57 species, although this is probably an underestimate of the true number. Additionally as discussed above, above co-kriging or indicator kriging may increase the number of species yet further.

6. REFERENCES

Chiles, J. & Delfiner, P. 1999. Geostatistics. Modeling Spatial Uncertainty. John Wiley & Sons, New York.

Gibbons, D.W., Reid, W.J.B. & Chapman, R.A. 1993. *The New Atlas of Breeding Birds in Britain and Ireland: 1988-91*. Poyser, London.

Gregory, R.D. & Baillie, S.R. 1998. Large-scale habitat use of some declining British birds. *Journal of Applied Ecology* **35:** 785-799.

Gregory, R.D., Bashford, R.I., Balmer, D.B., Marchant, J.H., Wilson, A.M. & Baillie, S.R. 1996. *The Breeding Bird Survey 1994-1995*. British Trust for Ornithology, Thetford, UK.

Haining, R. 1990. *Spatial data analysis in the social and environmental sciences*. Cambridge University Press.

Isaaks, E.H. & Srivastava, R.M. 1989. *An Introduction to Applied Geostatistics.* Oxford University Press, New York.

Johnston, K., Ver Hoef, J.M., Krivoruchko, K. & Lucas, N. 2001. *Using ArcGIS Geostatistical Analyst.* ESRI.

SAS. Institute Inc. 1996. SAS/Stat Software: Changes and Enhancements through Release 6.11. SAS Institute, Inc., Cary, North Carolina.

Acknowledgements

We are grateful to the thousands of volunteers who have contributed to the BTO/RSPB/JNCC Breeding Bird Survey (BBS), which provided the data on which this study was based. The BBS is jointly funded by the BTO, JNCC and the Royal Society for the Protection of Birds (RSPB). We thank Stephen Baillie for his comments on this report. Susan Waghorn provided much help with report production.

Table 1 A list of species in this study and subjective categorisation of their abundance and distribution in Britain.

Species	Abundance	Distribution
Feral Pigeon, *Columba livia*	A	P
Wood Pigeon, *Columba palumbus*	A	W
Collared Dove, *Streptopelia decaocto*	A	P
Ring-necked Parakeet, *Psittacula krameri*	R	L
Meadow Pipit, *Anthus pratensis*	A	P
Wren, *Troglodytes troglodytes*	A	W
Nightingale, *Luscinia megarhynchos*	R	L
Reed Warbler, *Acrocephalus scirpaceus*	A	L
Willow Tit, *Parus montanus*	A	L
Nuthatch, *Sitta europaea*	A	P
House Sparrow, *Passer domesticus*	A	W

Abundance: A - abundant, R - rare
Distribution: W - widespread, P - patchy, L - localised

Table 2 Comparison of the average standard error associated with predictions from the best prediction of abundance using manual diagnosis and modeling with prediction from a default model with parameters chosen by the program.

Species	Best prediction	Default model
Feral Pigeon	15.73	18.04
Wood Pigeon	19.60	19.71
Collared Dove	4.96	7.07
Ring-necked Parakeet	0.58	0.60
Meadow Pipit	4.79	8.78
Wren	10.16	11.25
Reed Warbler	1.41	1.46
Nuthatch	0.97	1.05
House Sparrow	20.25	28.01

Figure 1 Map showing the location of BBS squares surveyed in Britain in 2000

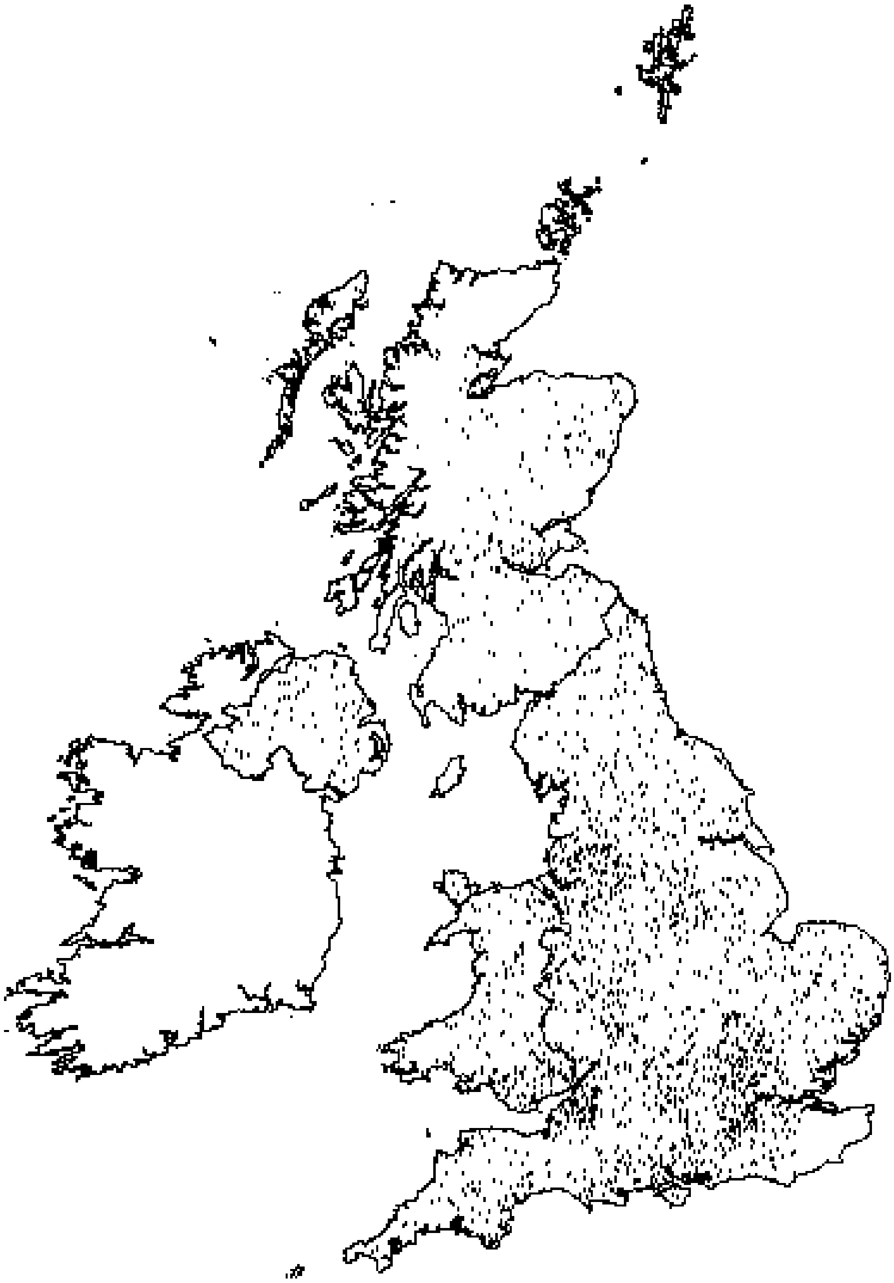
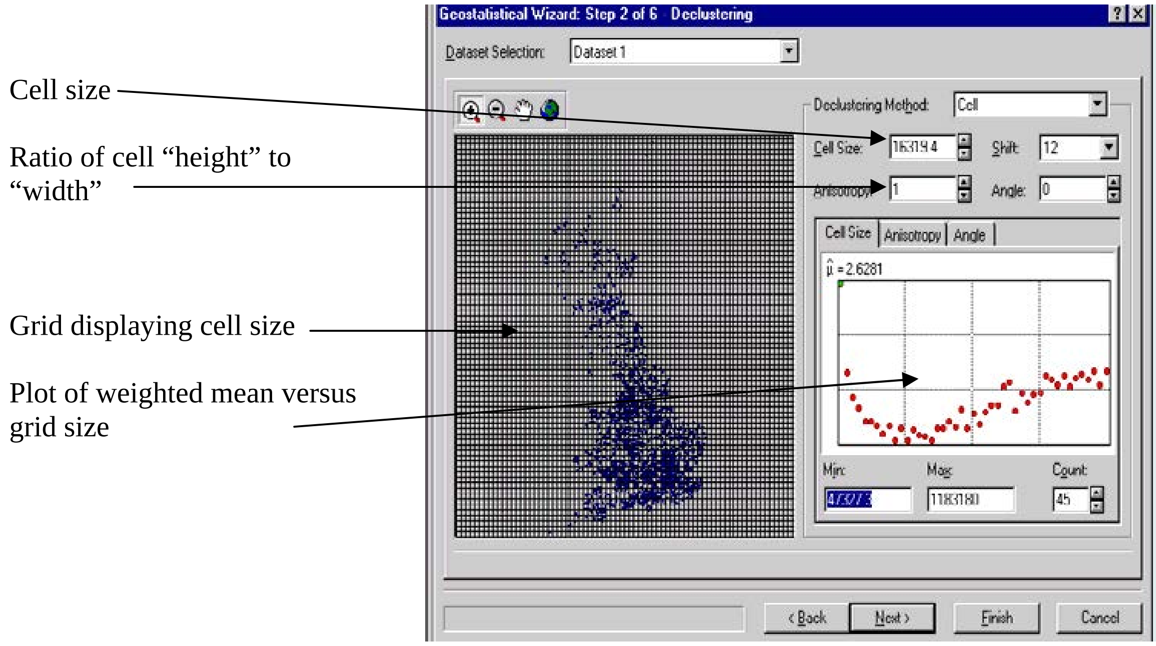


Figure 2 Declustering: to control for variation in observer coverage across Britain.

a. Cell declustering



b. Polygonal declustering

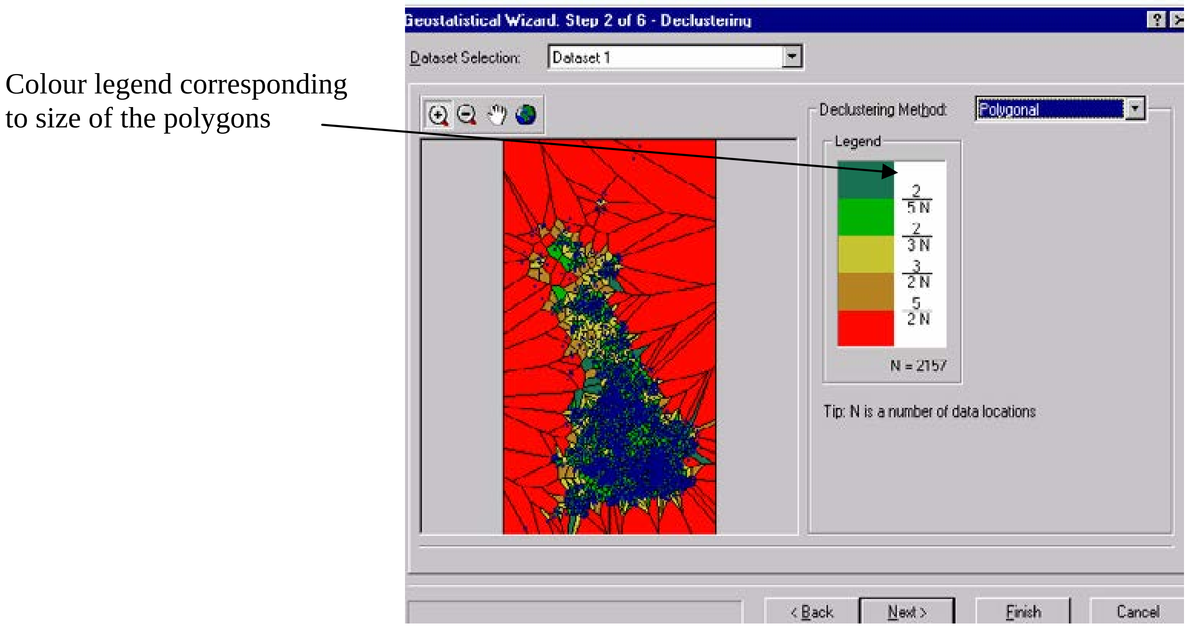


Figure 3 Projected abundance of Meadow Pipit in Britain

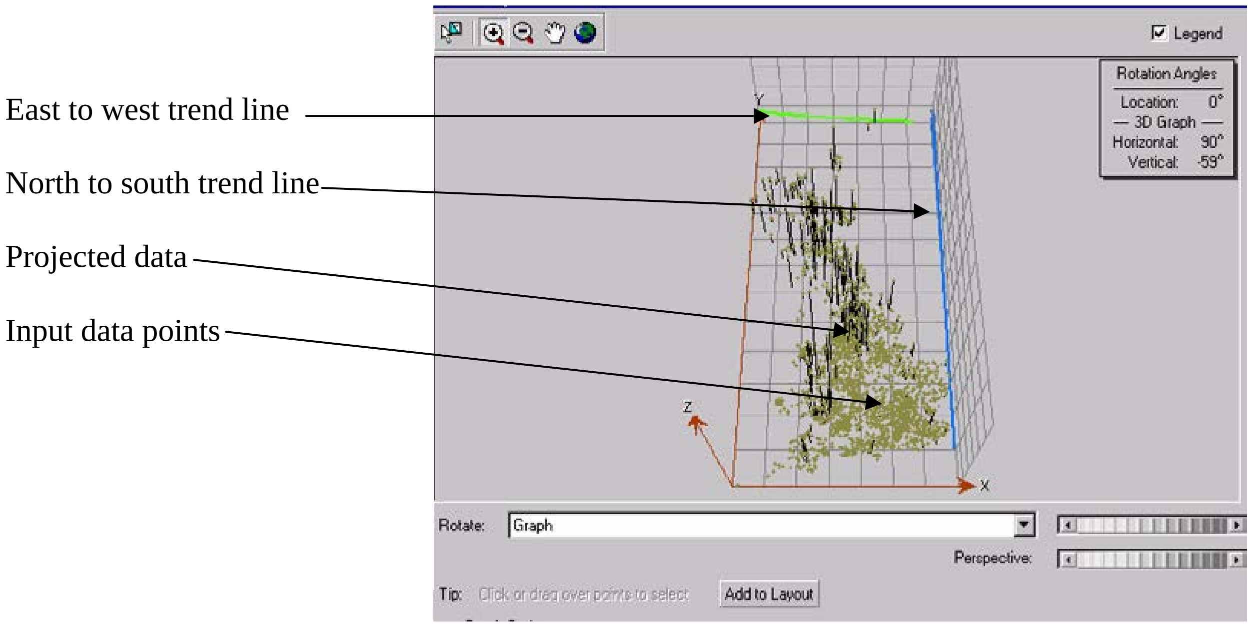


Figure 4 Semivariogram modelling in Geostatistical Analyst

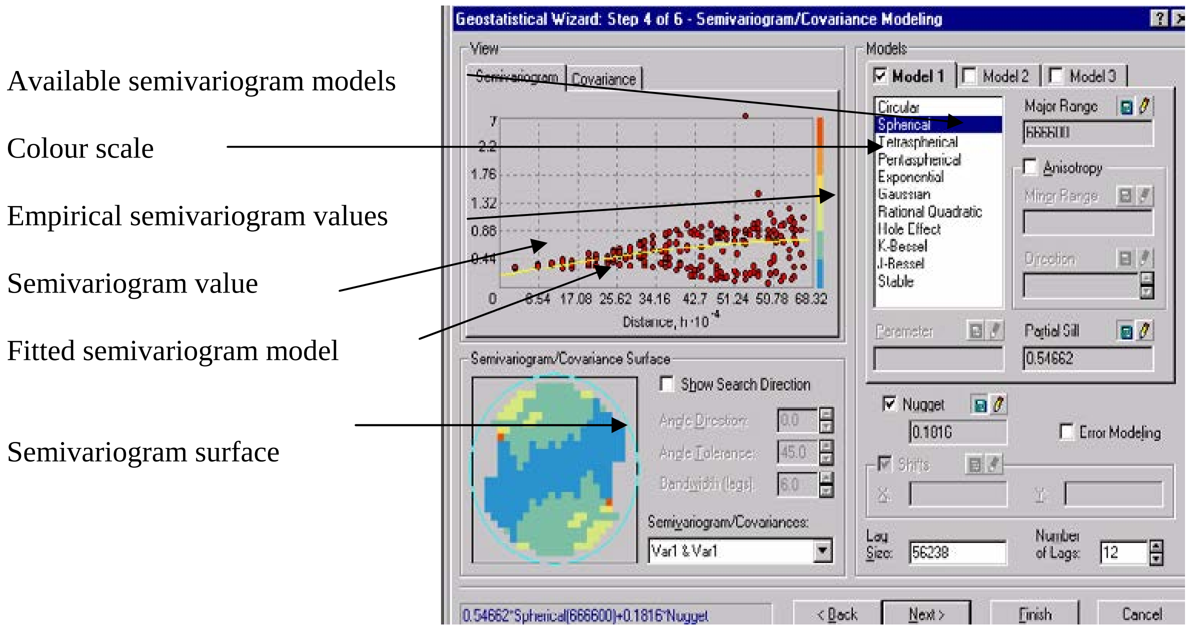
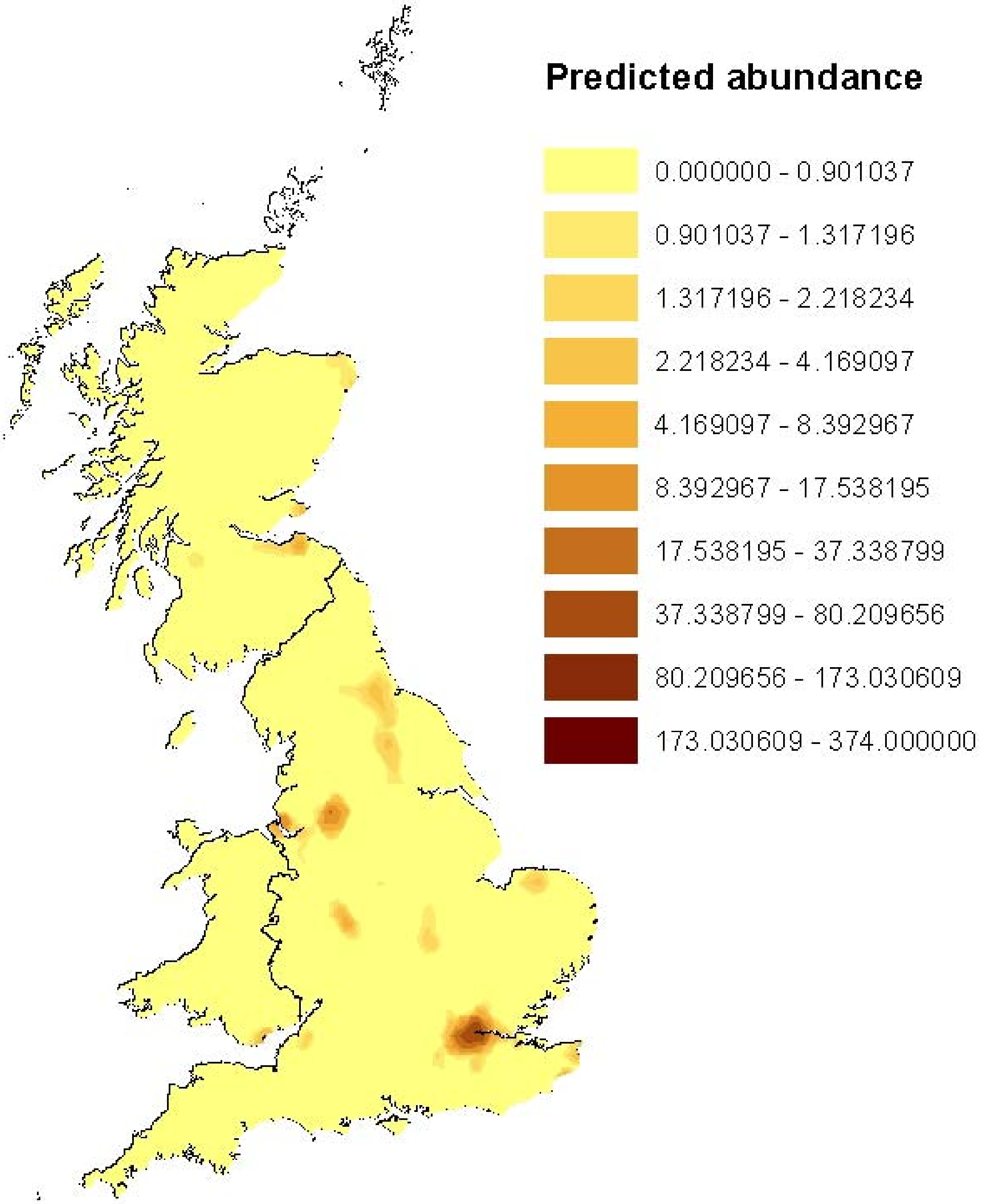
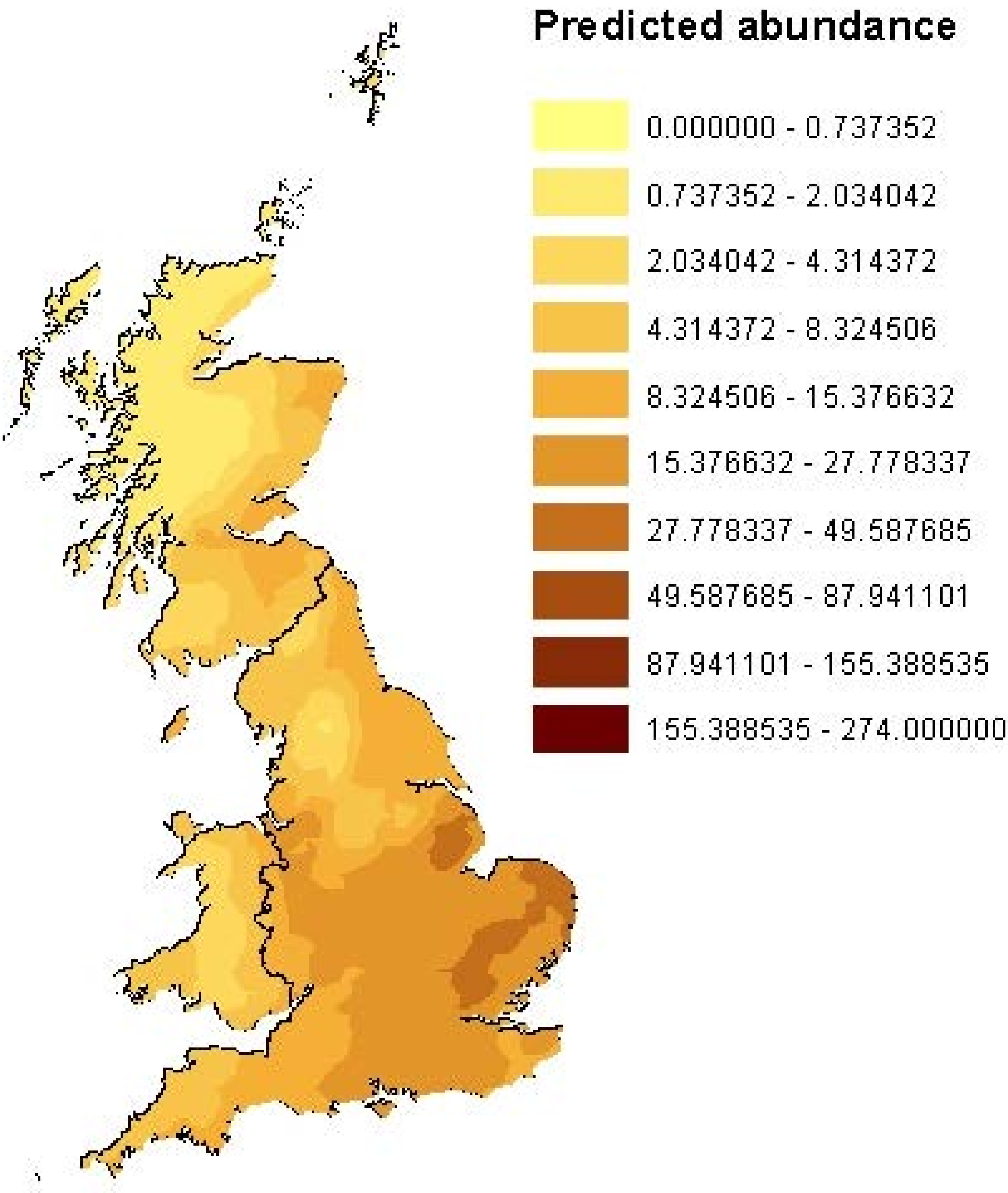


Figure 5 Maps of predicted abundance for nine bird species in Britain using BBS data for 2000.

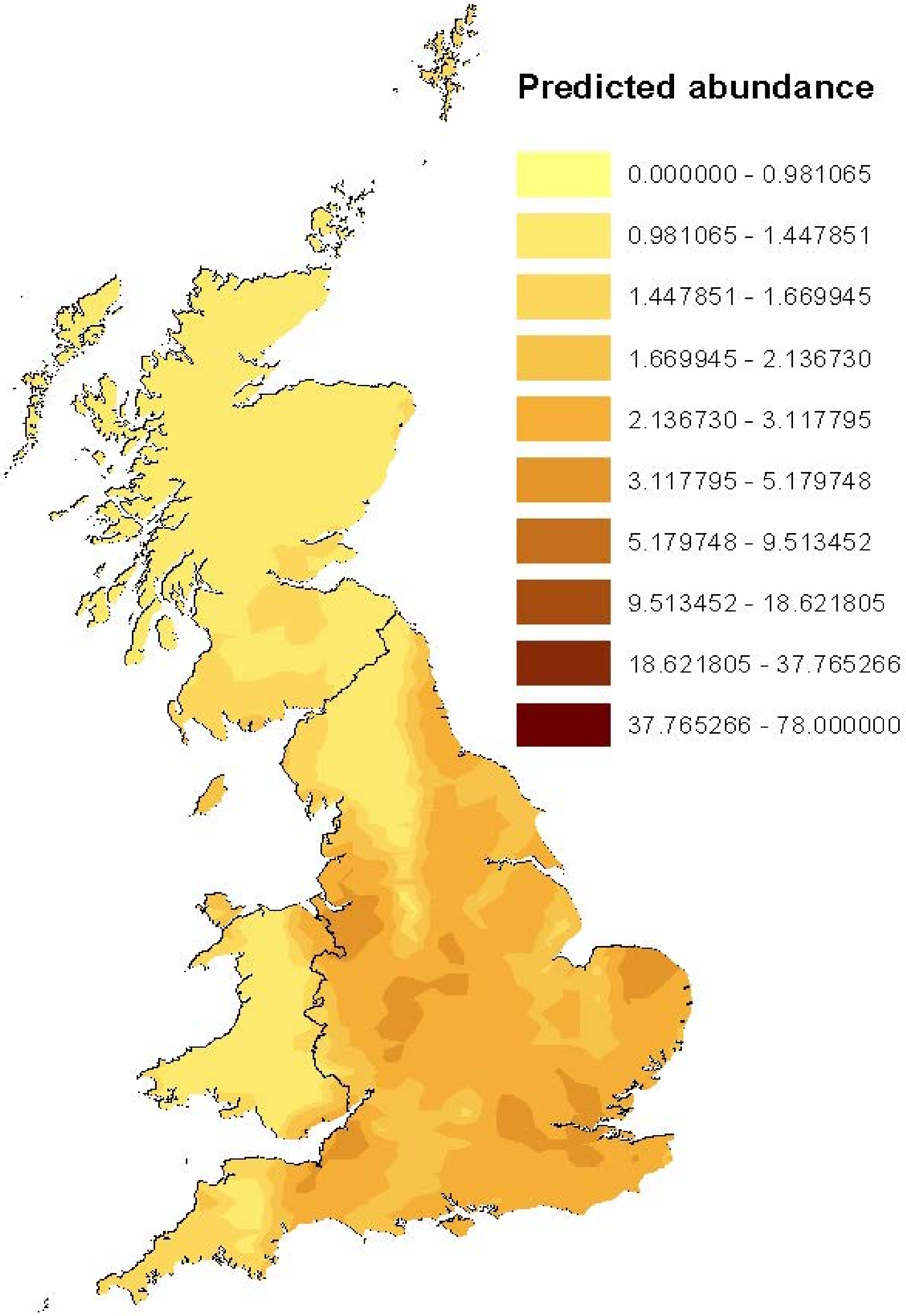
a. Feral Pigeon



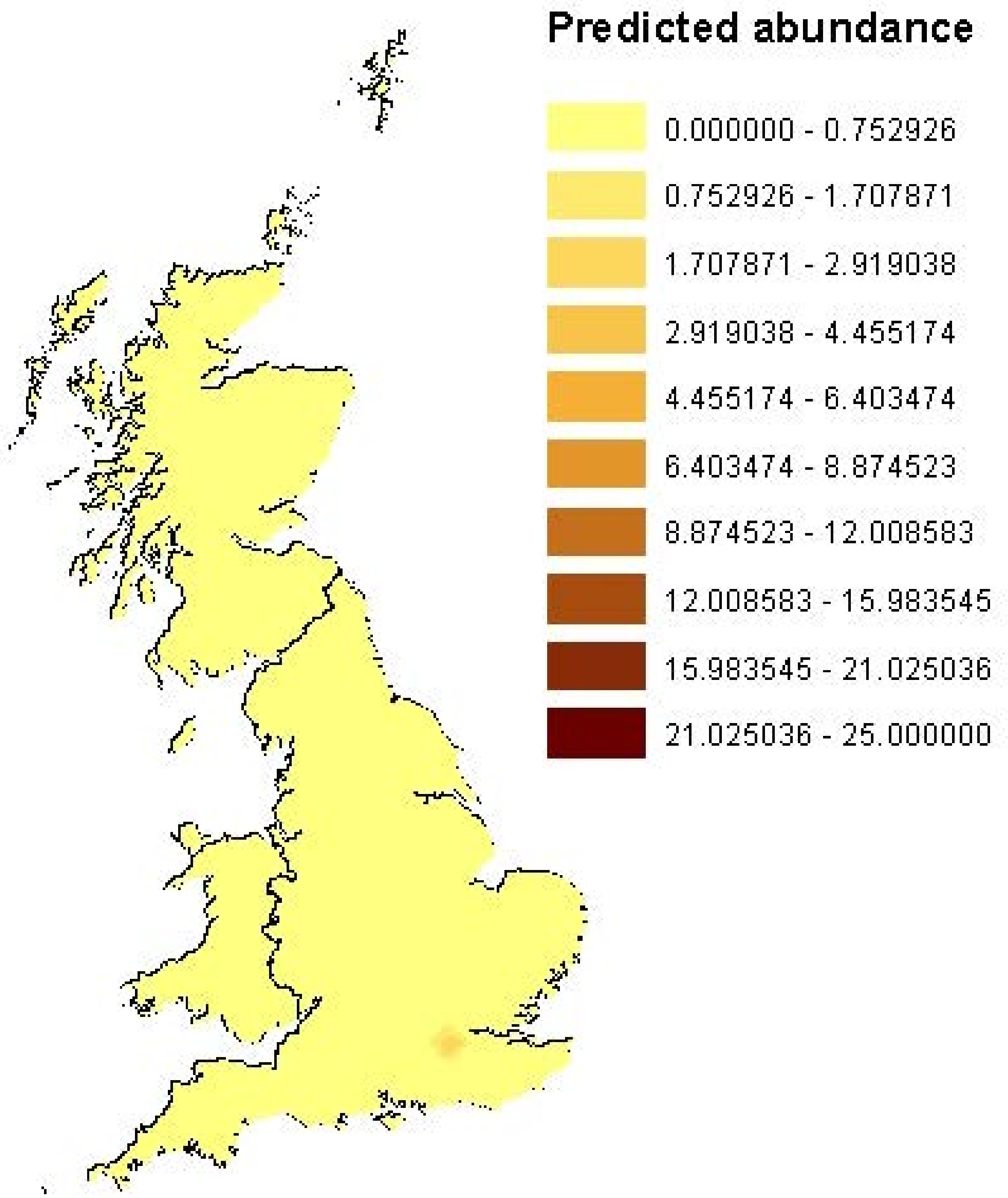
b. Wood Pigeon



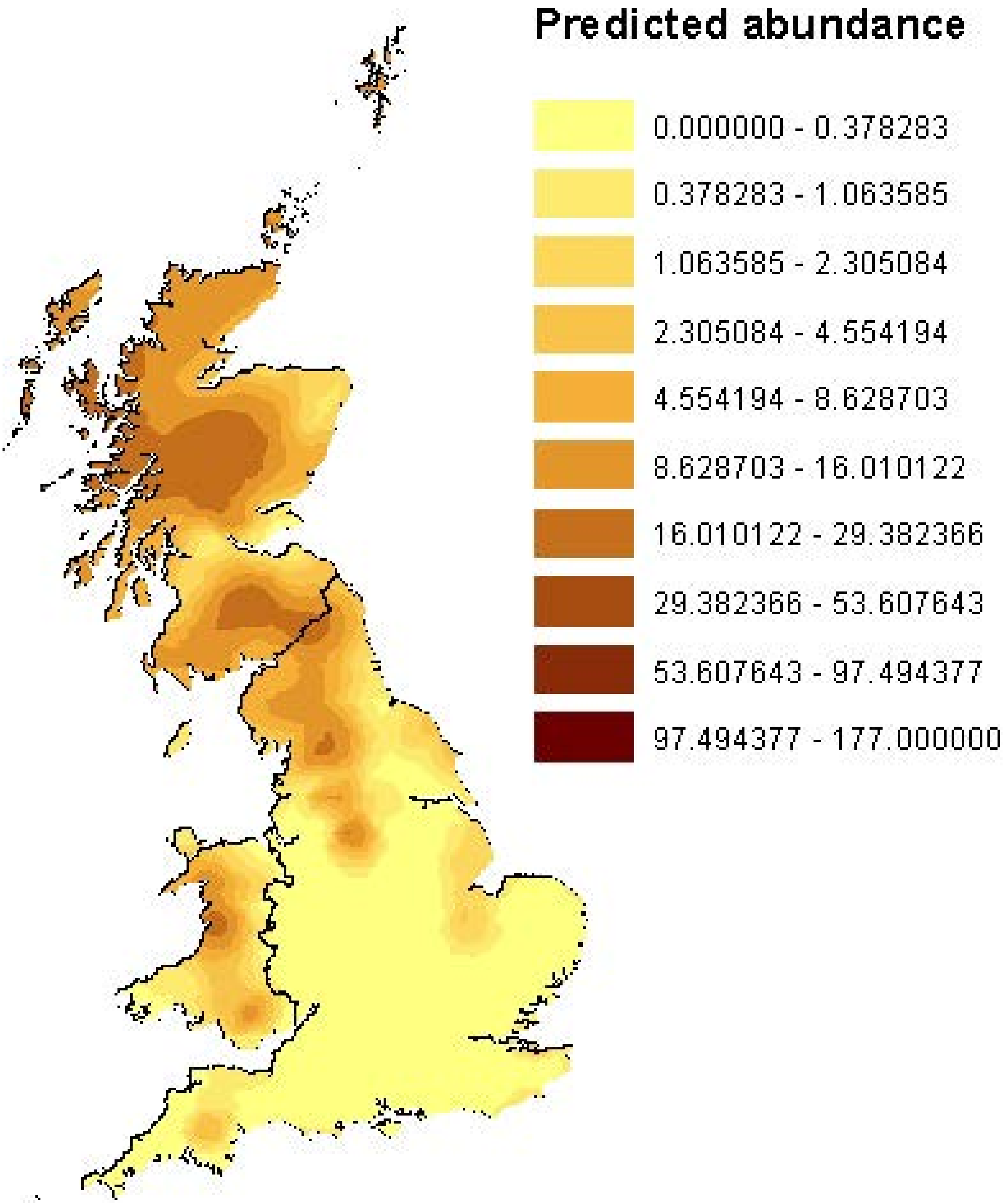
c. Collared Dove



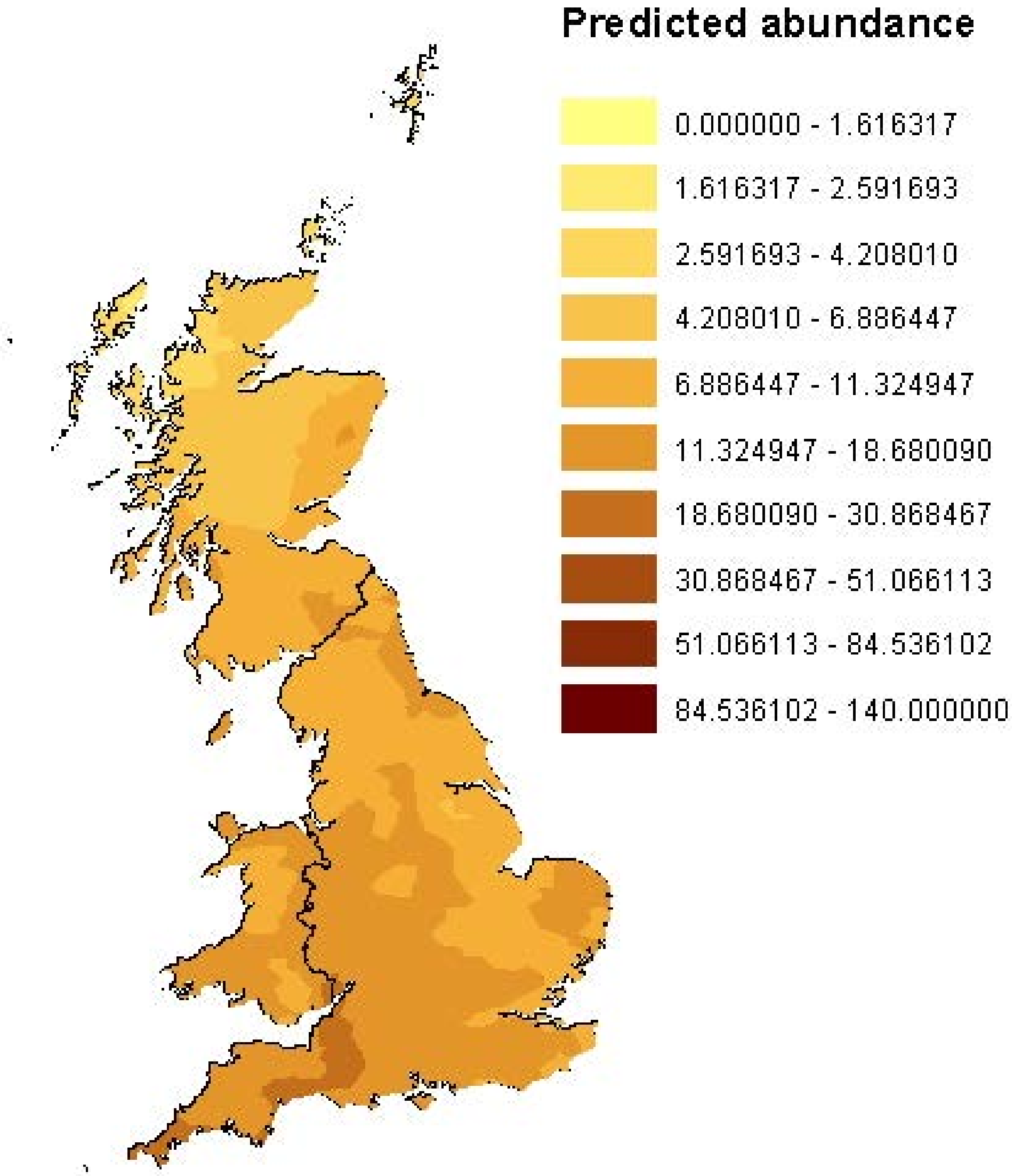
d. Ring-necked Parakeet



e. Meadow Pipit

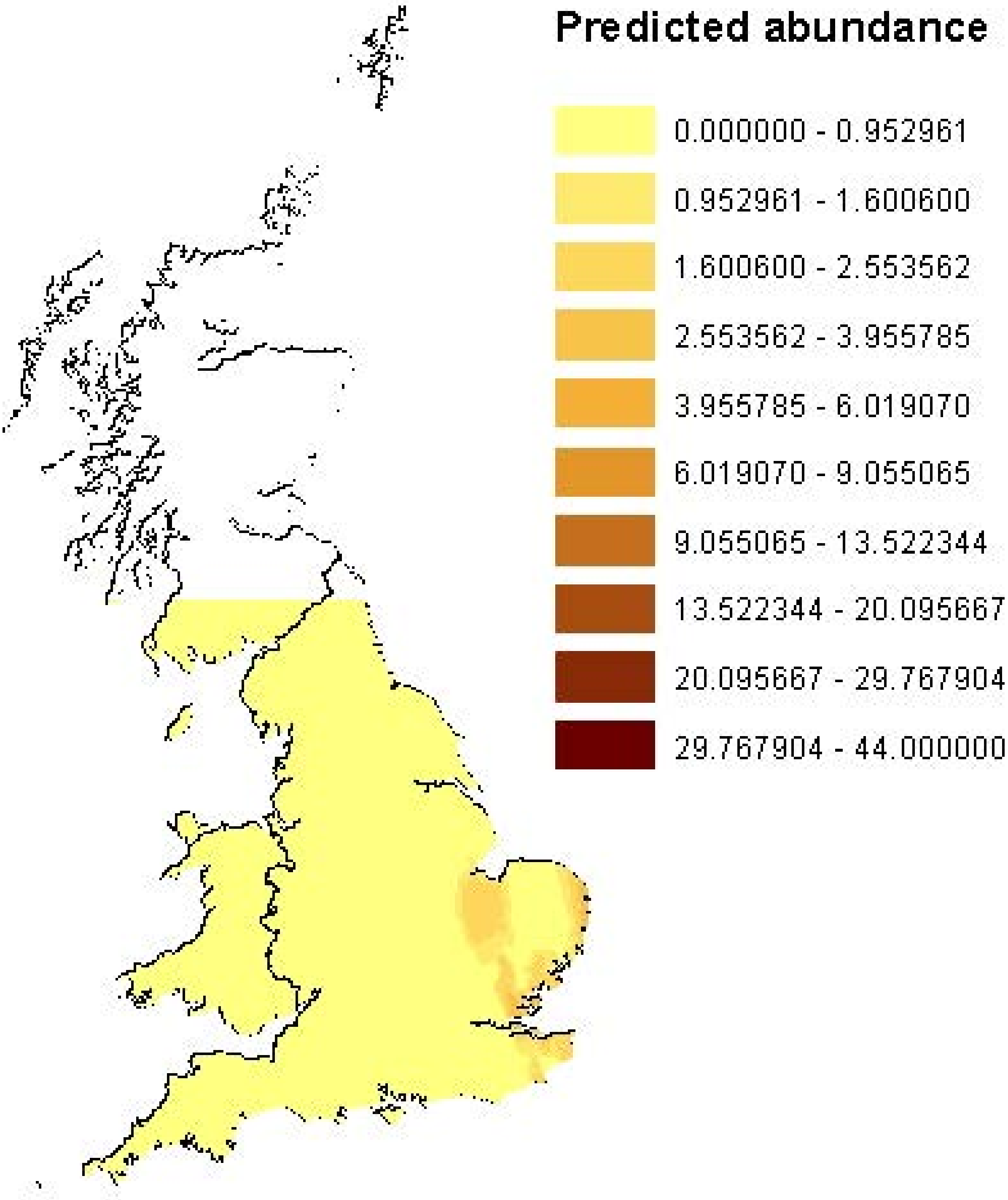


f. Wren

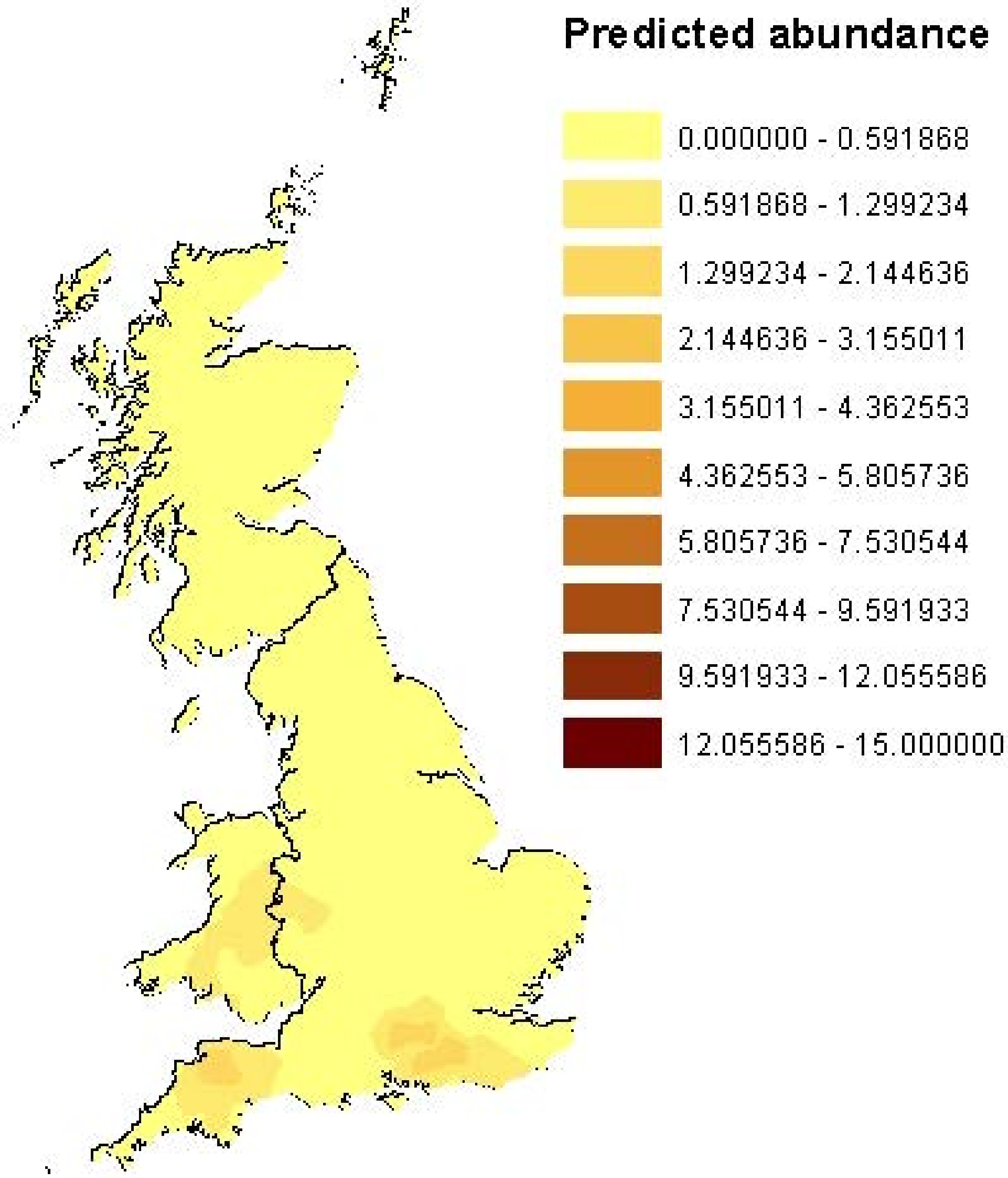


Predicted abundance
0.000000 - 1.616317
1.616317 - 2.591693
2.591693 - 4.208010
4.208010 - 6.886447
6.886447 - 11.324947
11.324947 - 18.680090
18.680090 - 30.868467
30.868467 - 51.066113
51.066113 - 84.536102
84.536102 - 140.000000

g. **Reed Warbler**



h. Nuthatch


Predicted abundance
0.000000 - 0.591868
0.591868 - 1.299234
1.299234 - 2.144636
2.144636 - 3.155011
3.155011 - 4.362553
4.362553 - 5.805736
5.805736 - 7.530544
7.530544 - 9.591933
9.591933 - 12.055586
12.055586 - 15.000000

i. **House Sparrow**

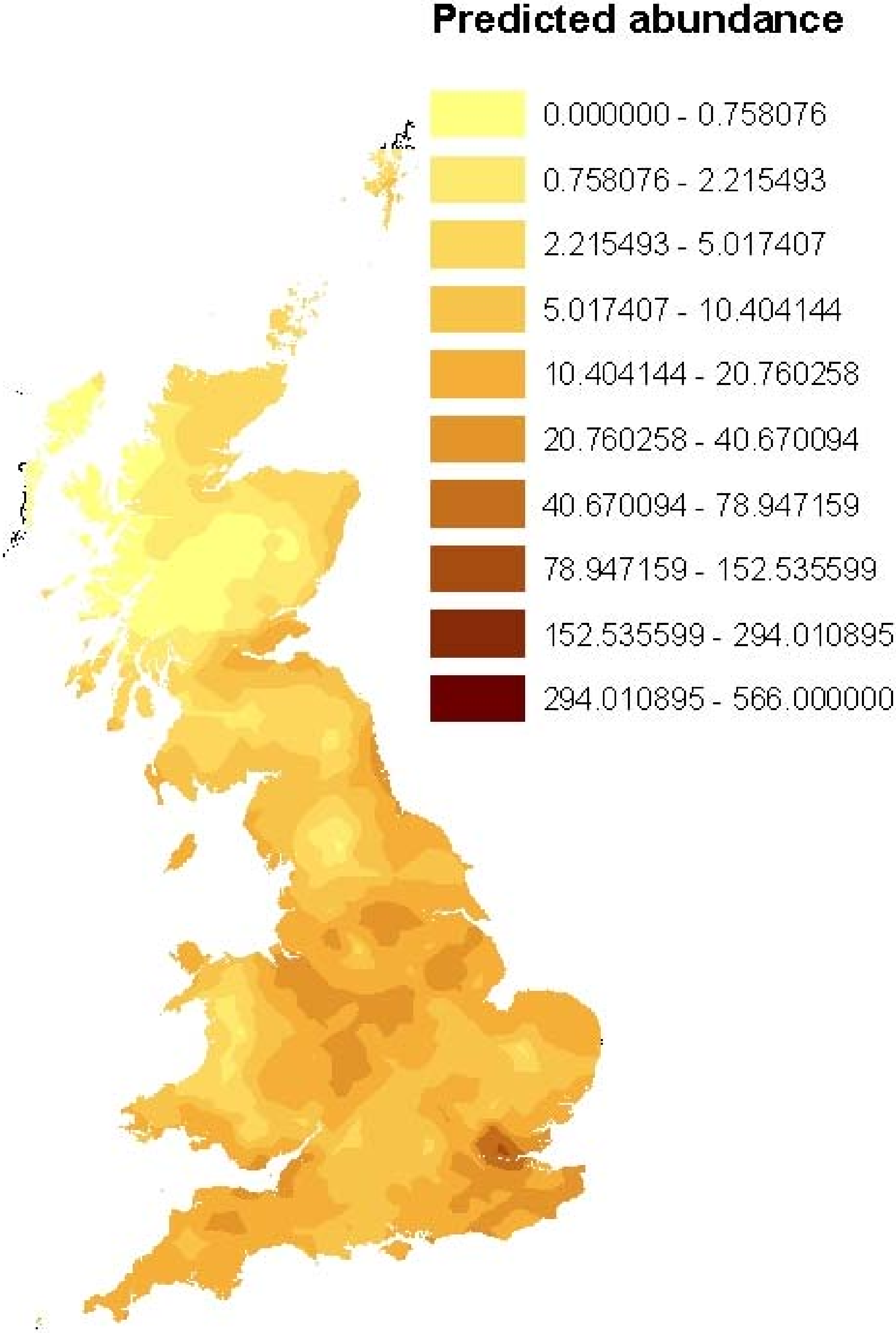
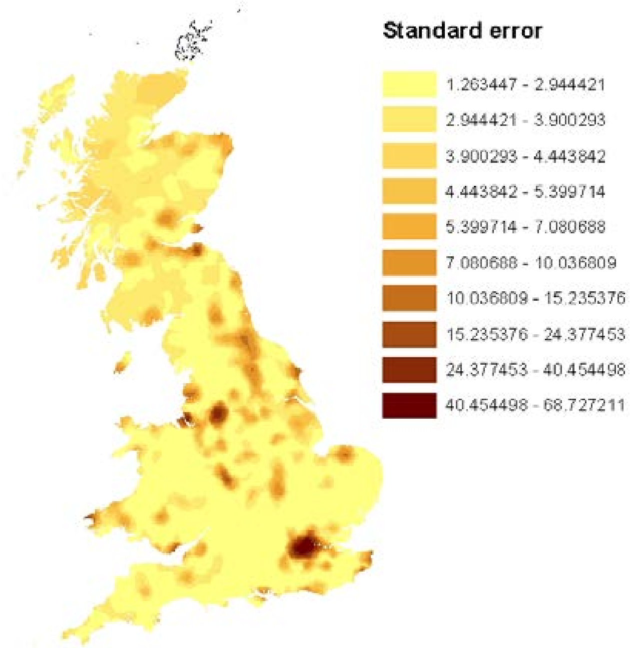
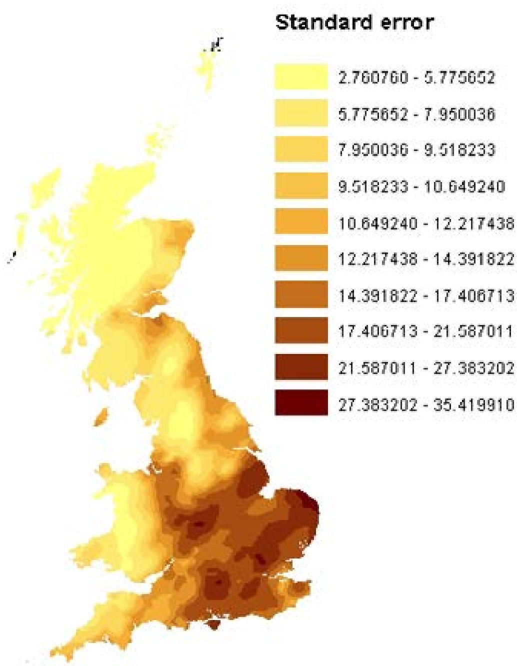


Figure 6 Maps of predicted standard error associated with abundance predictions for nine bird species in Britain using BBS data for 2000.

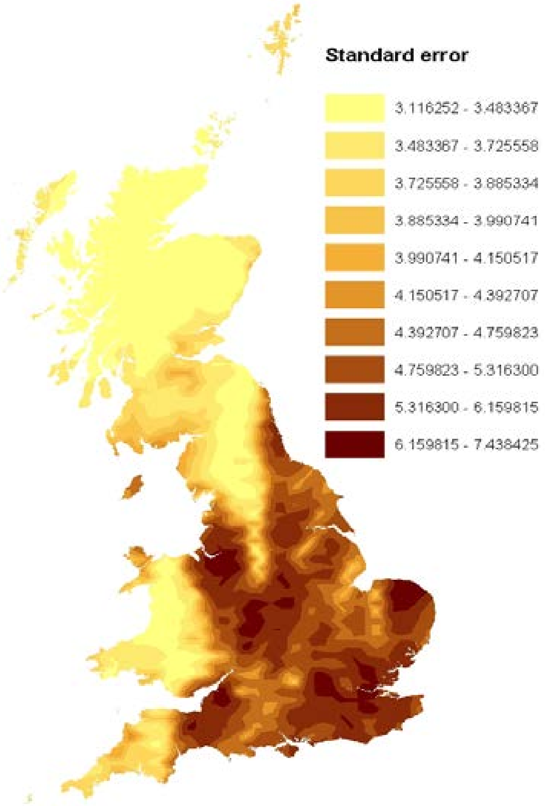
a. Feral Pigeon



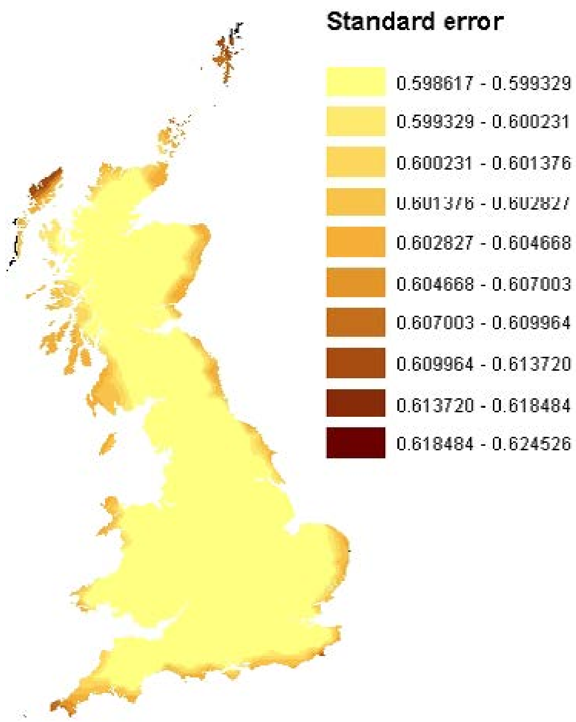
b. Wood Pigeon



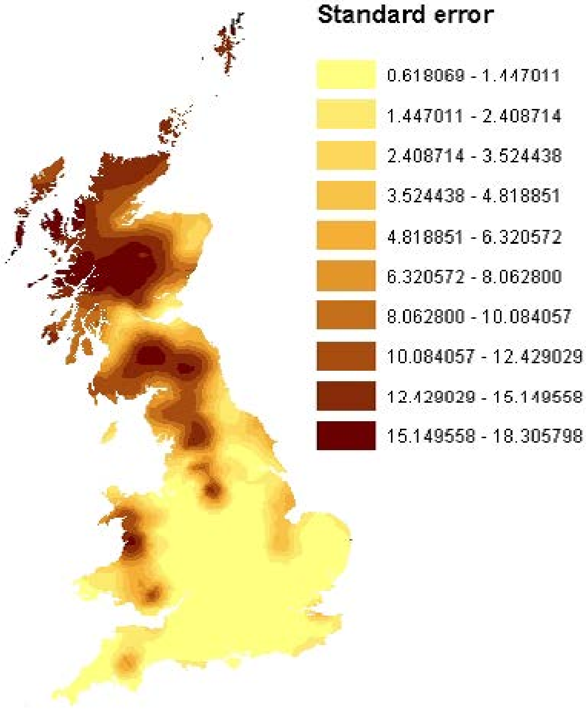
c. Collared Dove



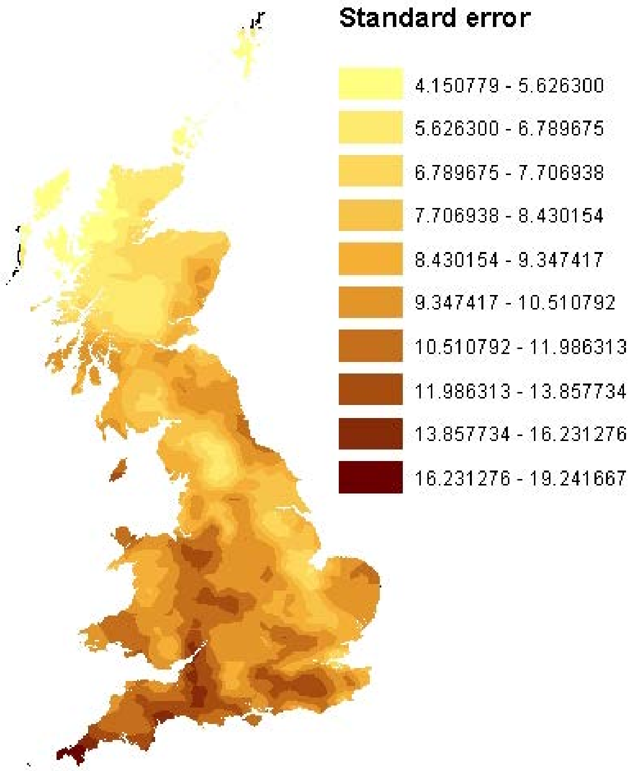
d. Ring-necked Parakeet



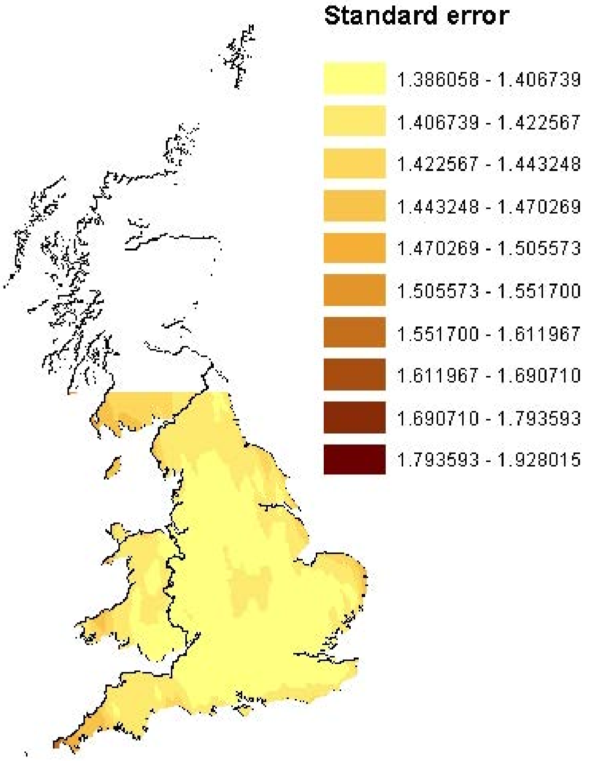
e. Meadow Pipit



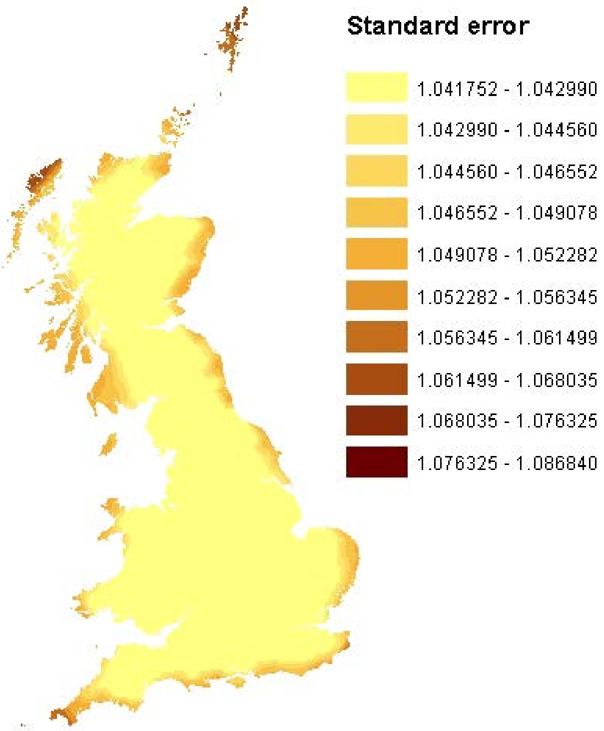
f. Wren



g. **Reed Warbler**



h. **Nuthatch**



i. House Sparrow

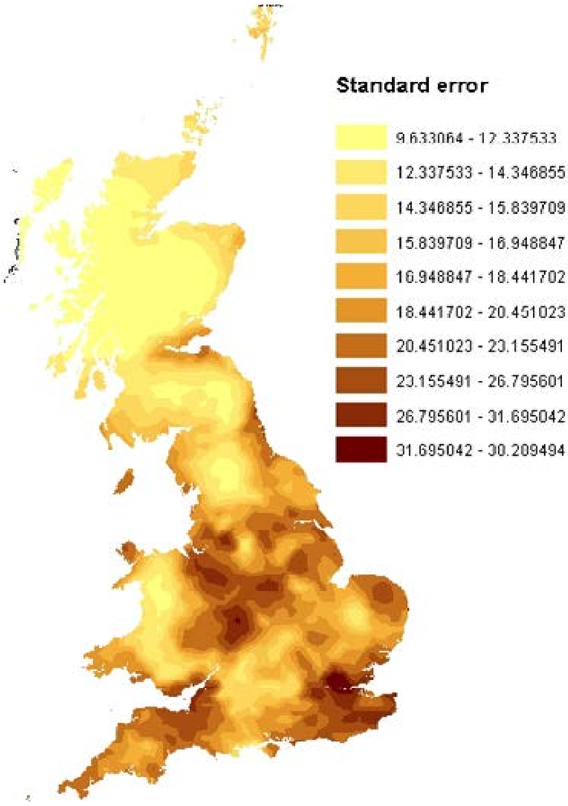
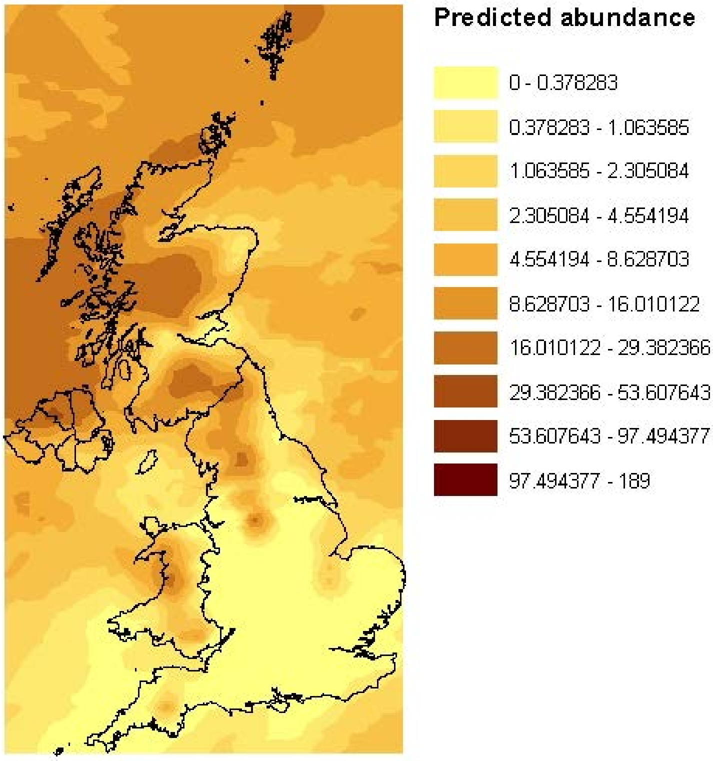


Figure 7 Maps of predicted abundance for two example bird species, Wren and Meadow Pipit in Britain and Northern Ireland using BBS data for 2000. We present predicted abundance across the range for interest to see how spatial autocorrelation from mainland Britain may contribute to predicted abundance in Northern Ireland.

a. Meadow Pipit



b. Wren

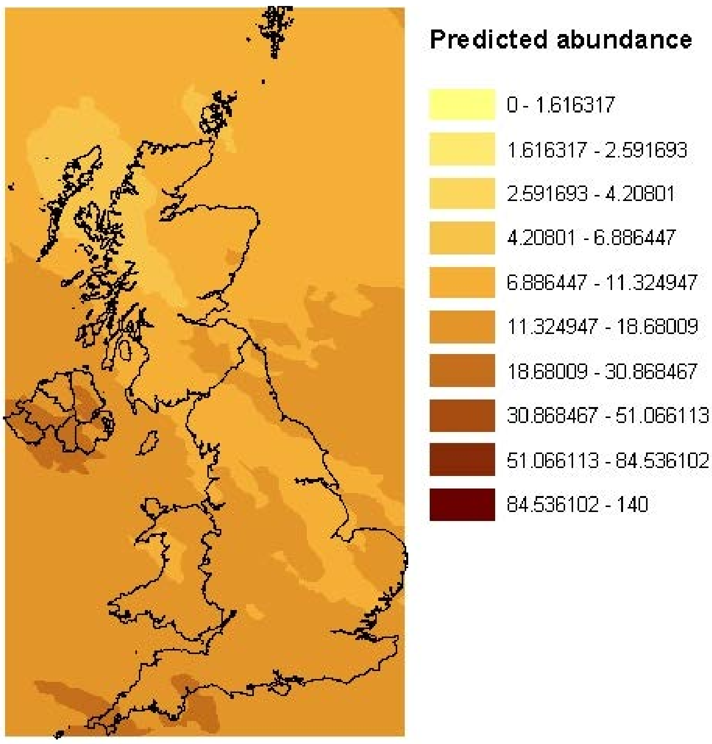
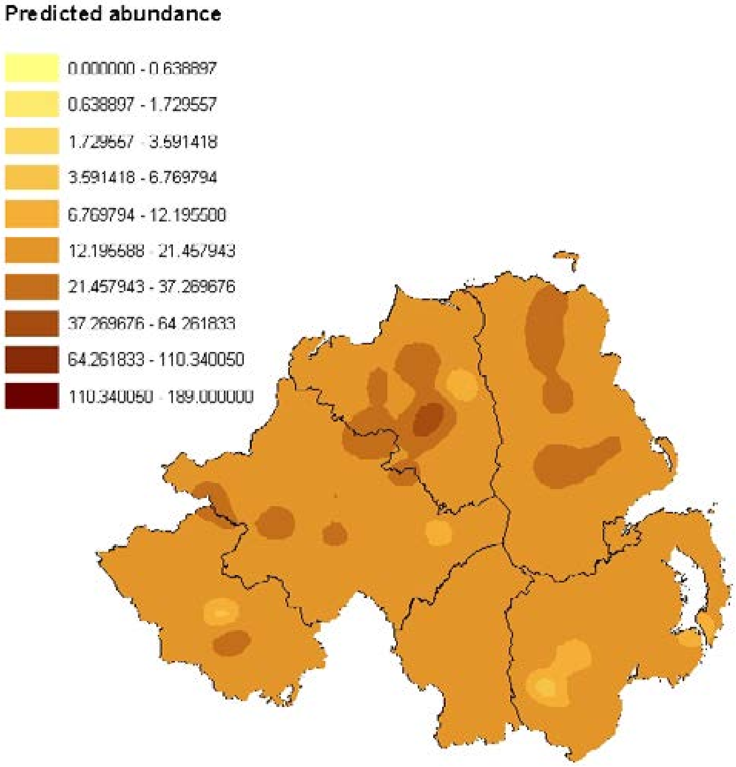


Figure 8 Maps of predicted abundance for two example bird species, Wren and Meadow Pipit in Northern Ireland using BBS data for 2000. We present predicted abundance across the range for interest to see how spatial autocorrelation from mainland Britain may contribute to predicted abundance in Northern Ireland. County boundaries are highlighted in a).

a. Meadow Pipit



b. Wren

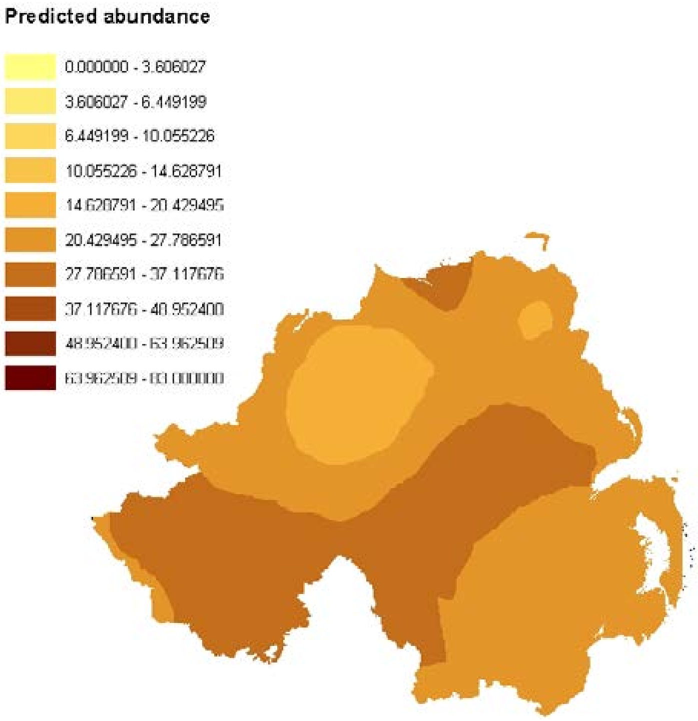
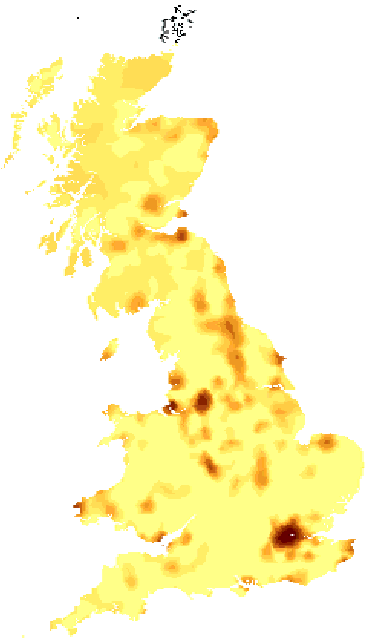


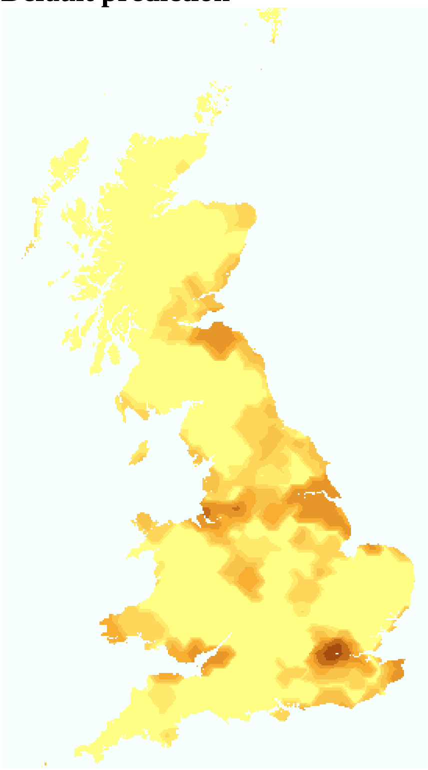
Figure 9 Comparison of best predictions from manual diagnoses and modeling of the data with predictions from a default model with parameters chosen by the program.

a. Feral Pigeon

Best prediction

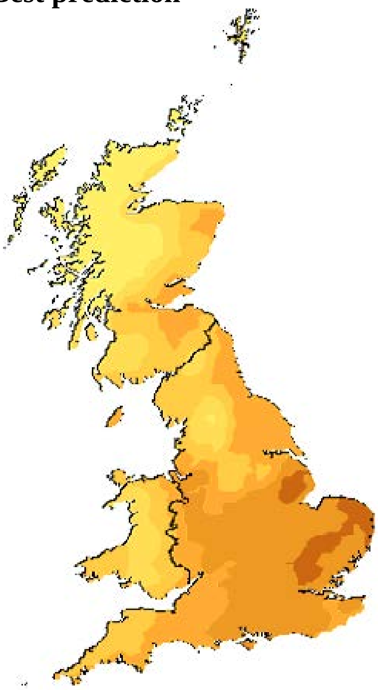


Default prediction

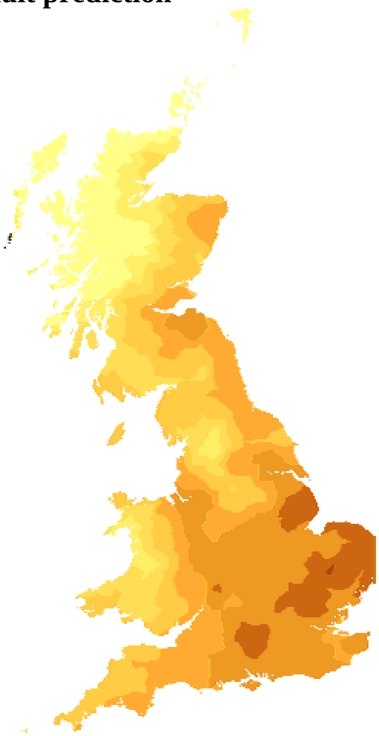


b. Wood Pigeon

Best prediction

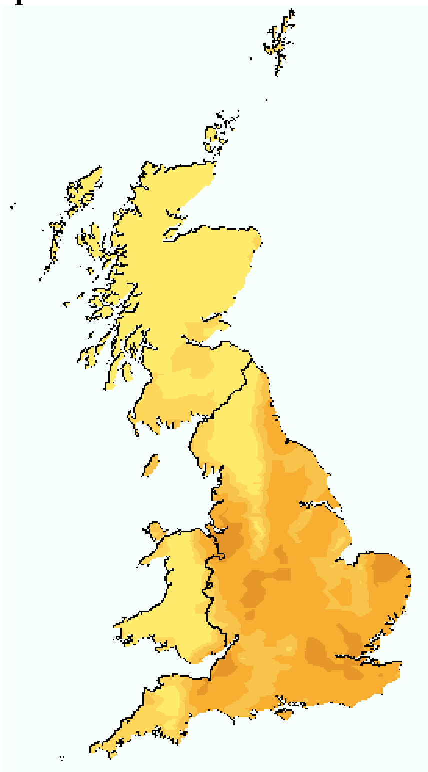


Default prediction

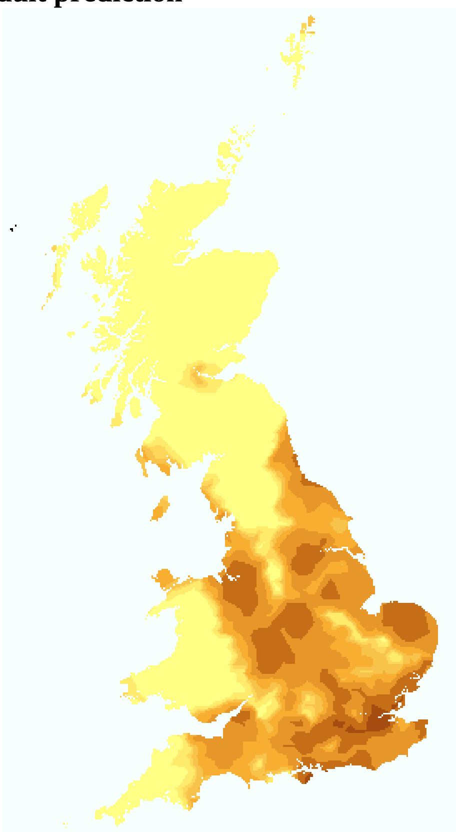


c. Collared Dove

Best prediction

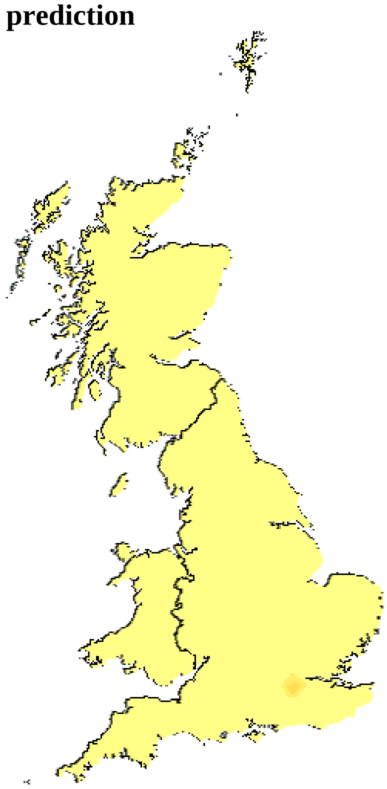


Default prediction

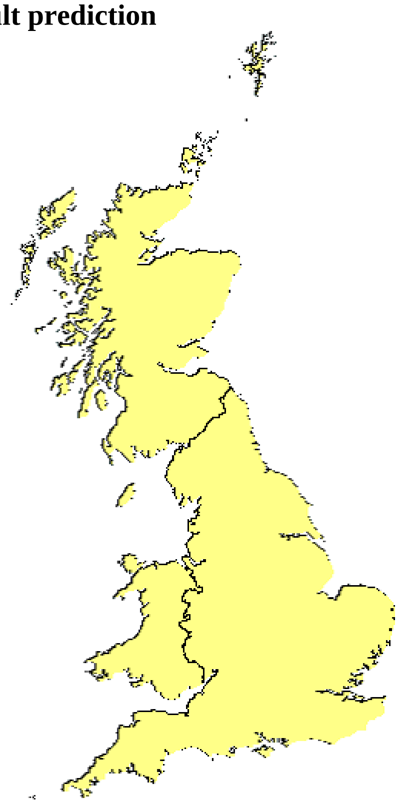


d. Ring-necked Parakeet

Best prediction

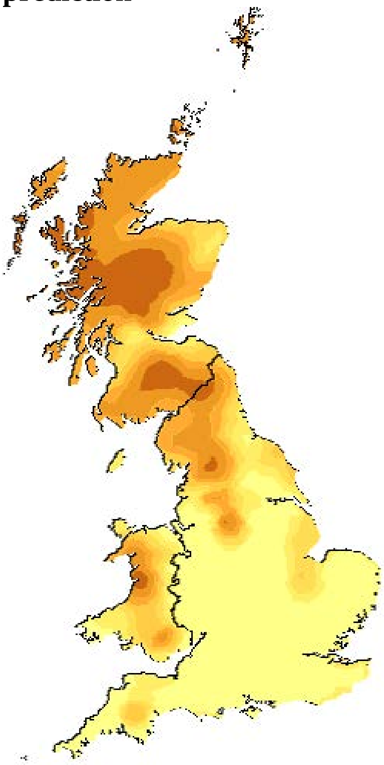


Default prediction

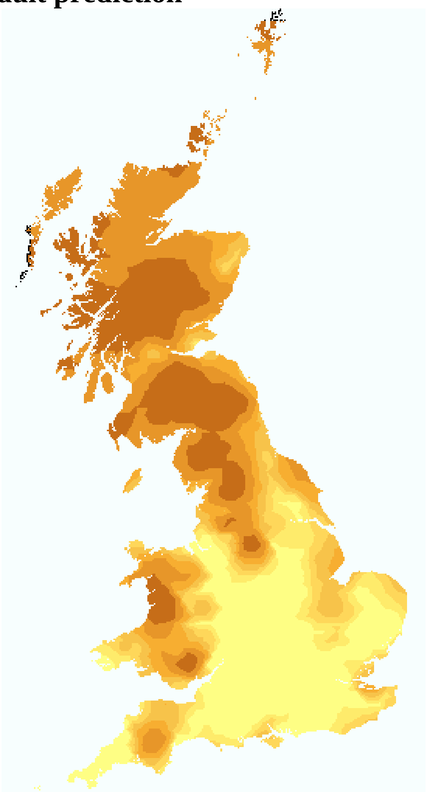


e) **Meadow Pipit**

Best prediction

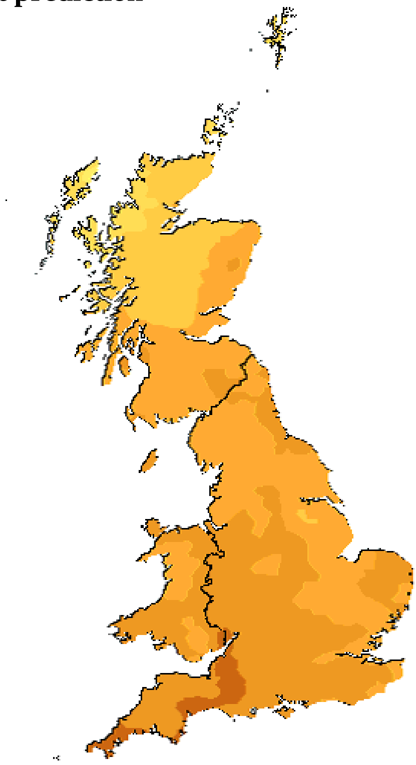


Default prediction

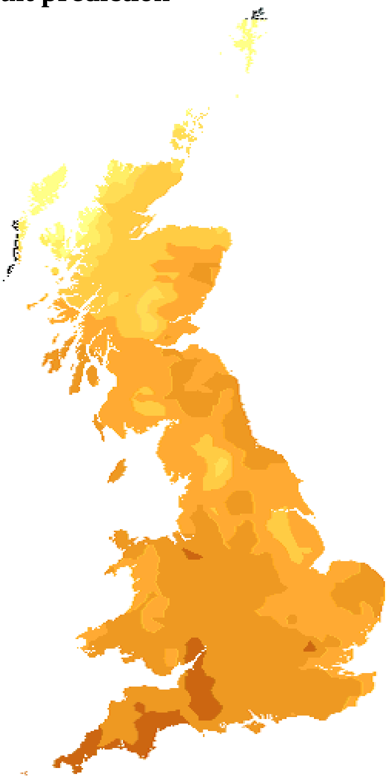


f. **Wren**

Best prediction

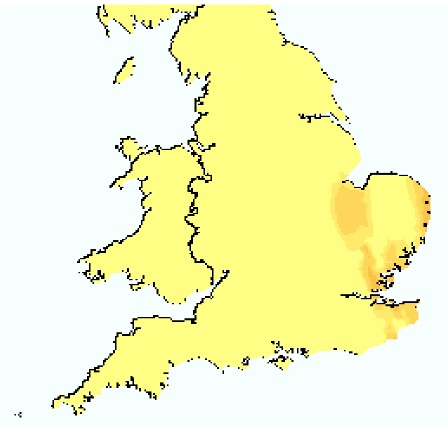


Default prediction

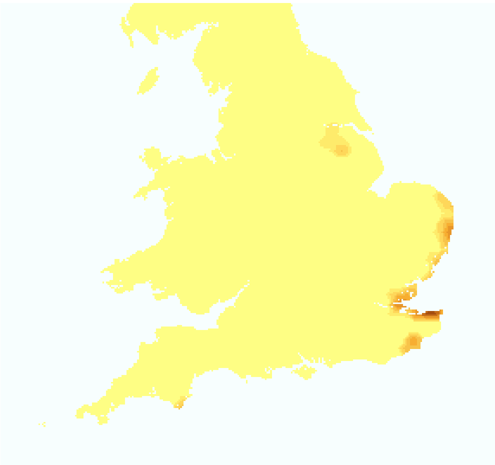


g) Reed Warbler

Best prediction

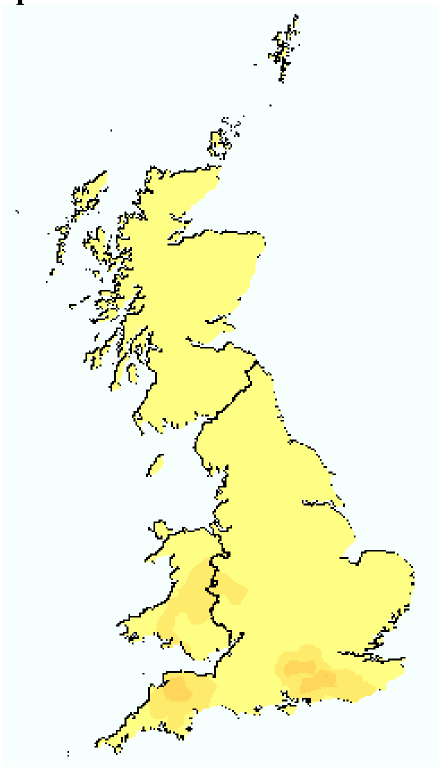


Default prediction

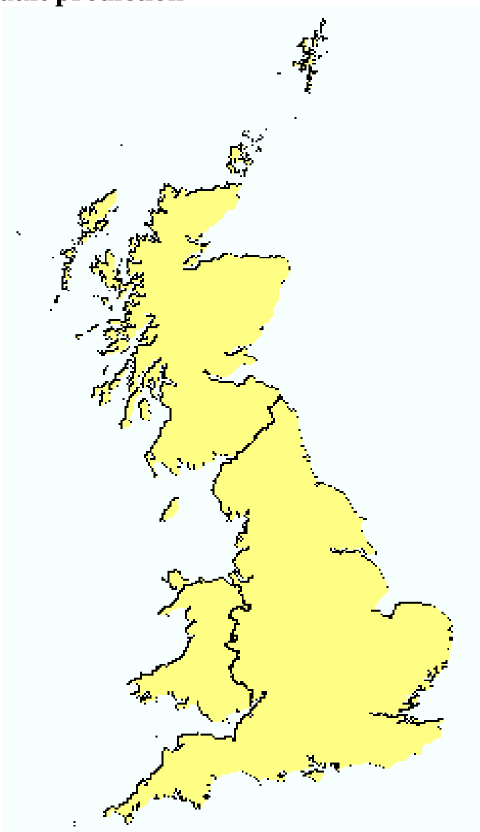


h) Nuthatch

Best prediction

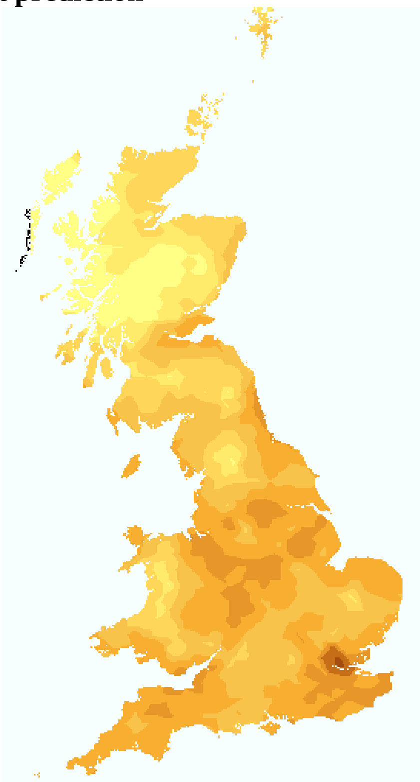


Default prediction



i) House Sparrow

Best prediction



Default prediction