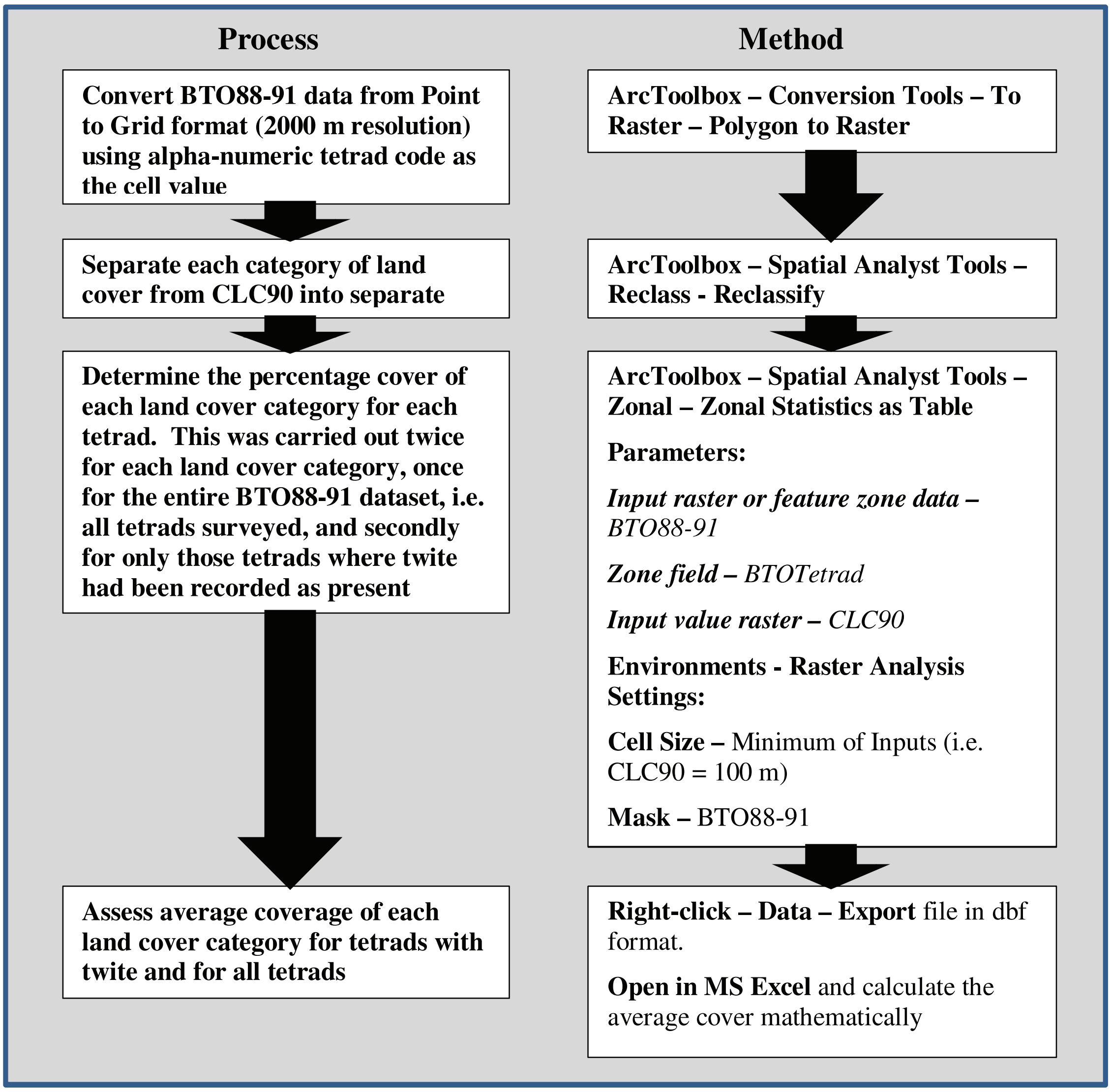
2.2 *Data Analysis*

A range of different forms of manipulation and analysis were carried out on the data described in the previous section in order to understand the relationships between different attributes and to develop the predictive models. Each form of analysis methods is described in the following sections.

2.2.1 Zonal Statistics as Table

The first step in the analysis process was to undertake some basic descriptive assessment of the land cover and agricultural census data and how they related to the distribution of twite. This was done using ArcGIS's ***ArcToolbox – Spatial Analyst Tools – Zonal – Zonal Statistics as Table*** tool (ESRI 2008). The process used for the CLC90 data is illustrated in Figure 15 below.



Process

Convert BTO88-91 data from Point to Grid format (2000 m resolution) using alpha-numeric tetrad code as the cell value

Separate each category of land cover from CLC90 into separate

Determine the percentage cover of each land cover category for each tetrad. This was carried out twice for each land cover category, once for the entire BTO88-91 dataset, i.e. all tetrads surveyed, and secondly for only those tetrads where twite had been recorded as present

Assess average coverage of each land cover category for tetrads with twite and for all tetrads

Method

ArcToolbox – Conversion Tools – To Raster – Polygon to Raster

ArcToolbox – Spatial Analyst Tools – Reclass - Reclassify

ArcToolbox – Spatial Analyst Tools – Zonal – Zonal Statistics as Table

Parameters:

Input raster or feature zone data – *BTO88-91*

Zone field – *BTOTetrad*

Input value raster – *CLC90*

Environments - Raster Analysis Settings:

Cell Size – Minimum of Inputs (i.e. CLC90 = 100 m)

Mask – BTO88-91

Right-click – Data – Export file in dbf format.

Open in MS Excel and calculate the average cover mathematically

Figure 15 – Example of the Zonal Statistics as Table procedure for CLC90 the data

This procedure was repeated for each land cover category in the CLC90, CLC20000 and LCS88 data sets and for the AgCensus data for 1988 and 2000. The results were assessed and, for all surveyed tetrads, the data for the top five land cover categories in twite tetrads and the four AgCensus attributes were consolidated into a new MS Excel table for each time period, i.e. CLC90_AgC88; CLC2000_AgC99; and LCS88_AgC88. The results of this analysis are presented and discussed in section 3.2.

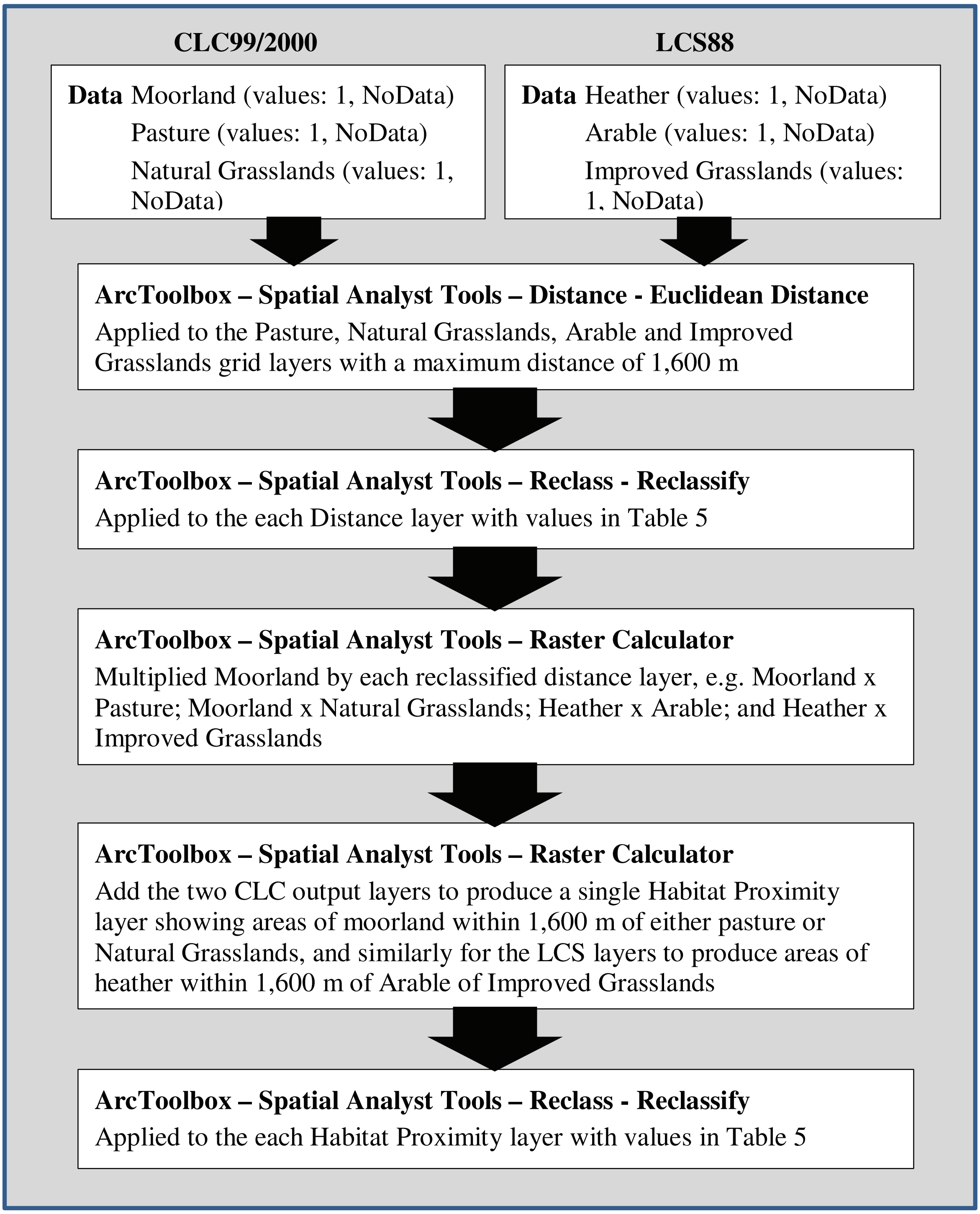
2.2.2 Habitat Proximity Estimation

While relatively simple statistics and data, such as the average for each land cover and land use category, might be useful in deriving a predictive model for twite distribution, the literature suggests there might be a more complex relationship between the birds and their environment (Orford 1973, Raine 2006, Wilkinson and Wilson 2010). Evidence suggests that birds tend to be found breeding on heather dominated moorland within 1.6 km of enclosed agricultural fields. Wilkinson and Wilson (2010) recorded the probability of finding birds within 400 m zones radiating out from enclosed farmland. In order to try and build a representation of this pattern into the predictive models an estimation of Habitat Proximity was derived following the procedure shown in Figure 16. The limitations of the available data were a critical factor in this analysis. The land cover and AgCensus data did not have any specific category for enclosed farmland. However, an estimate of agricultural land was made using the Pastures and the Natural Grasslands categories in the Corine data, and using the Arable and Improved Grasslands categories in the LCS data. The procedure shown in Figure 16 produced a new land cover moorland category, Habitat Proximity, which consisted of probabilities of twite being found corresponding to the distance of moorland to the above estimations of enclosed farmland, derived from Wilkinson and Wilson's observations in Uist and Harris, as shown in Table 5.

Habitat Proximity Value	Distance of Moorland to Enclosed farmland
0 %	Greater than 1,600 m
1.9 %	1,201 – 1,600 m
5.6 %	801 – 1,200 m
20.4 %	401 – 800 m
72.1 %	Up to 400 m

Table 5– Moorland Habitat Proximity values and distance to enclosed farmland

For this analysis it is critical that the original land cover layers have a value of '1' for cells with the land cover category, and 'NoData' for all cells without the category. Thus, when the layers are combined the product only has values for cells where there is an overlap between the moorland or heather and the appropriate estimate of enclosed farmland.



CLC99/2000

Data Moorland (values: 1, NoData)
Pasture (values: 1, NoData)
Natural Grasslands (values: 1, NoData)

LCS88

Data Heather (values: 1, NoData)
Arable (values: 1, NoData)
Improved Grasslands (values: 1, NoData)

ArcToolbox – Spatial Analyst Tools – Distance - Euclidean Distance
Applied to the Pasture, Natural Grasslands, Arable and Improved Grasslands grid layers with a maximum distance of 1,600 m

ArcToolbox – Spatial Analyst Tools – Reclass - Reclassify
Applied to the each Distance layer with values in Table 5

ArcToolbox – Spatial Analyst Tools – Raster Calculator
Multiplied Moorland by each reclassified distance layer, e.g. Moorland x Pasture; Moorland x Natural Grasslands; Heather x Arable; and Heather x Improved Grasslands

ArcToolbox – Spatial Analyst Tools – Raster Calculator
Add the two CLC output layers to produce a single Habitat Proximity layer showing areas of moorland within 1,600 m of either pasture or Natural Grasslands, and similarly for the LCS layers to produce areas of heather within 1,600 m of Arable of Improved Grasslands

ArcToolbox – Spatial Analyst Tools – Reclass - Reclassify
Applied to the each Habitat Proximity layer with values in Table 5

Figure 16 – Habitat Proximity evaluation procedure

2.2.3 Model Development - Logistic Regression

The main objective of this project was to derive a predictive model for twite distribution in Britain. As reported earlier in section 1.1.6, many different approaches have been applied to creating models for the prediction of presence/absence of different species, some more complex than others. In this case a relatively simple approach was taken based initially on logistic regression analysis of the relationships between twite distribution and the range of environmental and land use factors described in the preceding sections.

Regression analysis is typically used to explore relationships between two or more parameters (Bryman and Cramer (2005), Fotheringham et al (2002), Field (2005). Simple linear regression is used to investigate the relationship between two variables, producing an estimate of one variable based on the value of the other. Multiple regression goes a stage further and utilises a variety of parameters to produce an estimated outcome based on the relationships between them (Field 2005). In essence a simple linear regression equation is a basic model of the relationships between the parameters. A regression equation takes the form shown in equation 2.1:

$$Y_i = (b_0 + b_1X_i) + \varepsilon_i \qquad (2.1)$$

where, Y_i is the predicted outcome, X_i is the i^{th} value of the predictor variable, b_1 is the gradient of a straight line that is fitted to the data, and b_0 is the intercept of that line with the y axis. The parameters b_0 and b_1 are termed regression coefficients. The ε_i represents a residual value equivalent to the difference between the actual value of the i^{th} variable and the predicted value (Field 2005). Multiple regression follows the same principles, except that more variables are used, equation 2.2:

$$Y_i = (b_0 + b_1X_1 + b_2X_2 + \ldots. + b_nX_n) + \varepsilon_i \qquad (2.2)$$

where, Y is the outcome variable, b_1 is the coefficient of the first predictor (X_1), b_2 is the coefficient of the second predictor (X_2), b_n is the coefficient of the n^{th} predictor (X_n).

Logistic regression is used to predict a categorical (usually dichotomous) outcome from a set of predictor variables (Wuensch 2009). In the case of logistic regression the equation is slightly different as illustrated in equation 2.3 below:

$$P(Y) = \frac{1}{1 + e^{-(b_0+b_1X_1+b_2X_2+\ldots+b_nX_n+\,\varepsilon_i)}} \qquad (2.3)$$

where, *P(Y)* is the probability of *Y* occurring, *e* is the base of natural logarithms, and the other parameters and coefficients follow the same pattern as linear regression. The logarithmic transformation used in equation 2.3 corrects for possible breach of the condition of linear regression that there must be a linear relationship between each of the parameters and the outcome, which is unlikely when the outcome is dichotomous (Field 2005). Equation 2.3 produces a value between 0 and 1, where a value near 0 equates with a very low chance of the outcome Y occurring, while a value near 1 means that Y is likely to occur.

Thus, logistic regression is used to predict which of two categories of outcome will arise depending on a set of parameters. Hence, this type of regression analysis lends itself to being used to predict whether or not a species of bird might be present or not in any given area, depending upon the characteristics of that area, as defined by a set of variables. This forms the basis of the following investigations into the relationships between twite distribution and the previously described environmental and land use variables.

A stepped process of logistic regression analysis was employed to develop a range of increasingly complex models aimed at predicting the presence of twite across Britain. Prior to undertaking the logistic regression analysis the three sets of land cover data were compared with the British/Scottish mean coverage for each land cover category and with the average coverage for tetrads with twite. The same was done for the AgCensus88 data. These comparisons were used to create new attributes (e.g. *Moor_Ave*) where each record was allocated a score of 1 if the value was above the British Average and 0 if it was below the British Average, except for Pasture where the relationship was negative so the inverse score was applied, i.e. 0 = above average, 1 = below average. Similarly, if the value was above or below the Tetrad Average (e.g. *Moor_TetAve*). No average scores were applied to the Conifers category as there was no obvious difference between tetrads with or without twite. The various data sets were prepared for this analysis by joining them to the BTO88 twite tetrad data by their spatial location using the ArcGIS ***Right-click – Joins & Relates – Join…*** function and setting the parameters to ***Join data from another layer based on spatial location***. This produced a new data set with the fields describe in Table 6 below.

Attribute	Description	Format
BTO_Tetrad	Alpha-numeric tetrad code	Text
Easting	British National Grid x coordinate	Integer
Northing	British National Grid y coordinate	Integer
P_A	Presence or absence of twite (0 = absence, 1 = presence)	Integer
Moor	% cover of moorland per tetrad	Float
Moor_Ave	Above or below British average cover of moorland (0 = below, 1 = above)	Integer
Moor_TetAve	Above or below Tetrad average cover of moorland (0 = below, 1 = above)	Integer
Pasture	% cover of pasture per tetrad	Float
Pasture_Ave	Above or below British average cover of pasture (0 = above, 1 = below)	Integer
Pasture_TetAve	Above or below Tetrad average cover of pasture (0 = above, 1 = below)	Integer
NGrass	% cover of natural grassland per tetrad	Float
NGrass_Ave	Above or below British average cover of natural grassland (0 = below, 1 = above)	Integer
NGrass_TetAve	Above or below Tetrad average cover of natural grassland (0 = below, 1 = above)	Integer
Conifers	% cover of conifers per tetrad	Float
Peat	% cover of peat per tetrad	Float
Peat_Ave	Above or below British average cover of peat (0 = below, 1 = above)	Integer
Peat_TetAve	Above or below Tetrad average cover of peat(0 = below, 1 = above)	Integer
CLC_Ave	Sum of the CLC comparisons with British Average (0-4)	Integer
CLC_TetAve	Sum of the CLC comparisons with Tetrad Average (0-4)	Integer
Cattle	Number of cattle per tetrad	Float
Cattle_Ave	Above or below British average number of cattle (0 = below, 1 = above)	Integer
Cattle_TetAve	Above or below Tetrad average number of cattle (0 = below, 1 = above)	Integer
Sheep	Number of sheep per tetrad	Float
Sheep_Ave	Above or below British average number of sheep (0 = below, 1 = above)	Integer
Sheep_TetAve	Above or below Tetrad average number of sheep (0 = below, 1 = above)	Integer
Crop	Area of crops (ha per tetrad)	Float
Crop_Ave	Above or below British average area of crops (0 = below, 1 = above)	Integer
Crop_TetAve	Above or below Tetrad average area of crops (0 = below, 1 = above)	Integer
Fallow	Area of fallow (ha per tetrad)	Float
Fallow_Ave	Above or below British average area of fallow (0 = below, 1 = above)	Integer
Fallow_TetAve	Above or below Tetrad average area of fallow (0 = below, 1 = above)	Integer
AgC_Ave	Sum of the AgC comparisons with British Average (0-4)	Integer
AgC_TetAve	Sum of the AgC comparisons with Tetrad Average (0-4)	Integer
CLC_AgC_Ave	CLC_TetAve - AgC_TetAve (-4 to 4)	Integer
CLC_AgC_TetAve	CLC_TetAve - AgC_TetAve (-4 to 4)	Integer
Elevation	Elevation a.s.l. (m)	Integer
R_May	LTA Monthly Rainfall for May (mm)	Float
R_Jun	LTA Monthly Rainfall for June (mm)	Float
R_Jul	LTA Monthly Rainfall for July (mm)	Float
R_Aug	LTA Monthly Rainfall for August (mm)	Float
R_Sept	LTA Monthly Rainfall for September (mm)	Float
R_Ave	Average LTA Monthly Rainfall during the breeding season (mm)	Float
T_May	LTA Daily Temperature for May (^{o}C)	Float
T_Jun	LTA Daily Temperature for June (^{o}C)	Float
T_Jul	LTA Daily Temperature for July (^{o}C)	Float
T_Aug	LTA Daily Temperature for August (^{o}C)	Float
T_Sept	LTA Daily Temperature for September (^{o}C)	Float
T_Ave	Average LTA Daily Temperature during the breeding season (^{o}C)	Float

Table 6 – Attributes for Logistic Regression Model Development

Each data table from the joined layers was exported as a dbf table and opened in PASW Statistics v.18 (previously SPSS) for the logistic regression analysis. Initially a correlation analysis was undertaken for all the variables against each other to establish the nature and extent of any relationships, and to identify any collinearity between variables. There were some obvious relationships, e.g. between *T_Ave* and each of the monthly average temperature variables. For subsequent analysis care was taken to ensure that variables with clear collinearity were not included in the same logistic regression analysis.

A simple logistic regression was carried out on all variables against the *P_A* variable, using the **ENTER** method. This produced a score for each variable that reflected the strength of its relationship with *P_A*. Using the scores as a guide a range of combinations was used for logistic regression analysis, using the **ENTER** method, and saving the predicted probability of a 'twite present' outcome (*PRE*), the predicted group membership (*PGR*), i.e. 1 or 0 (twite present or not) based on a cut-off value of ≥50% probability for presence, and the standardised residual value (the difference between the predicted probability and the actual value (*ZRE*). The output of the various attempts at logistic regression were saved as an MS Excel file and a new cut-off value was calculated that produced the same number of tetrads with twite as the original dataset. In all cases the cut-off value was well below the 50% threshold. This was undoubtedly due to the fact that in the raw data there was a great many more tetrads without twite than with twite, only 3.2% of the surveyed tetrads had twite present in them. This tends to lead to a bias in the regression analysis in favour of a negative outcome, and hence produces lower probabilities.

During the logistic regression analysis several measures of the 'goodness of fit' for the model were also calculated by the PASW software including: the -2 Log likelihood (-2LL); Chi-square; Cox & Snell R Square; and Nagelkerke R Square. These values are derived in different ways but all reflect how well the model predicts the actual data. In addition to these standard measures a further ROC analysis and calculation of the Akaike Information Criterion (AIC) were carried out to determine whether or not the models we working well. A summary of these measures is presented at Appendix 5.

A summary of the steps followed to undertake and interpret the logistic regression analyses are presented in Figure 17.

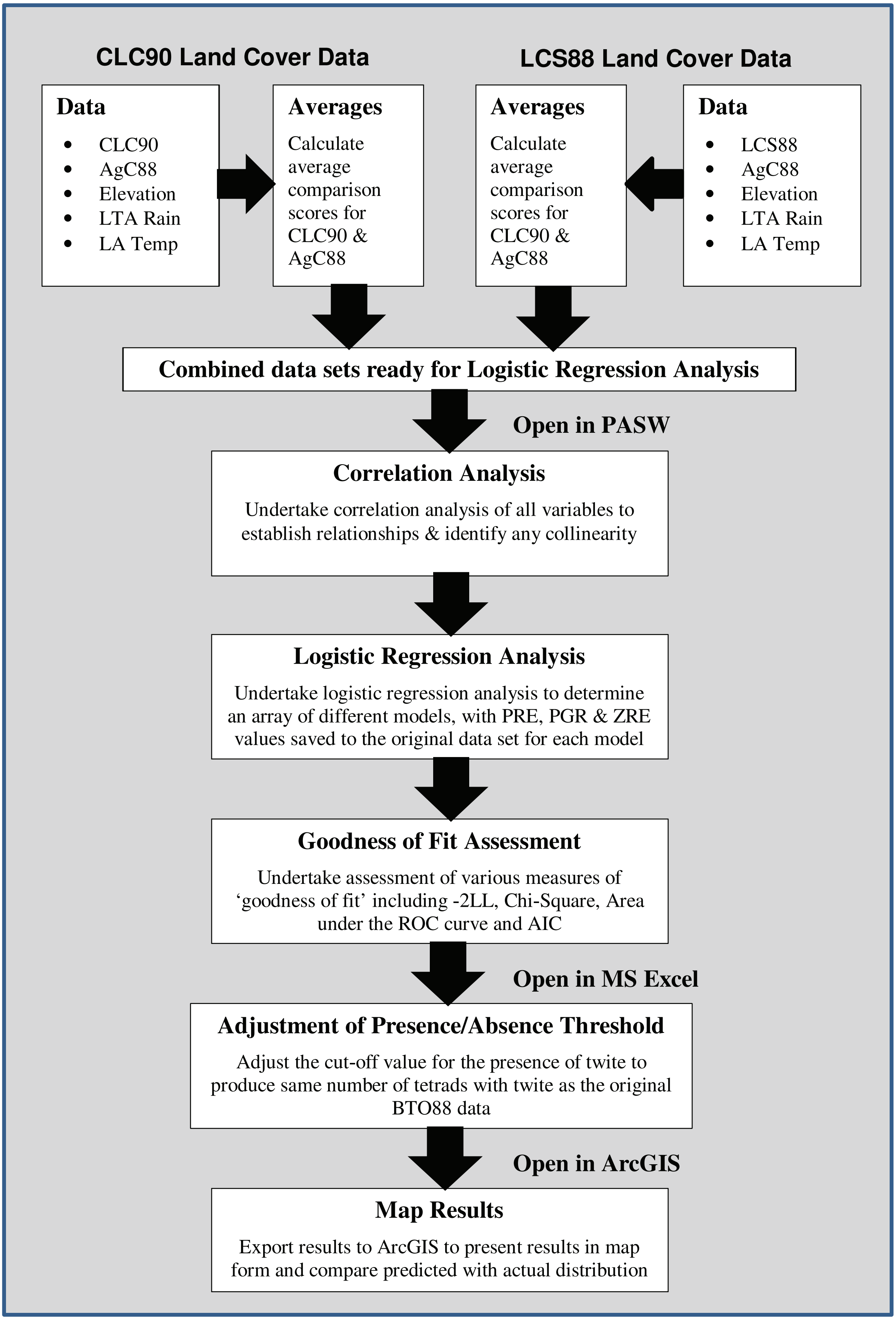
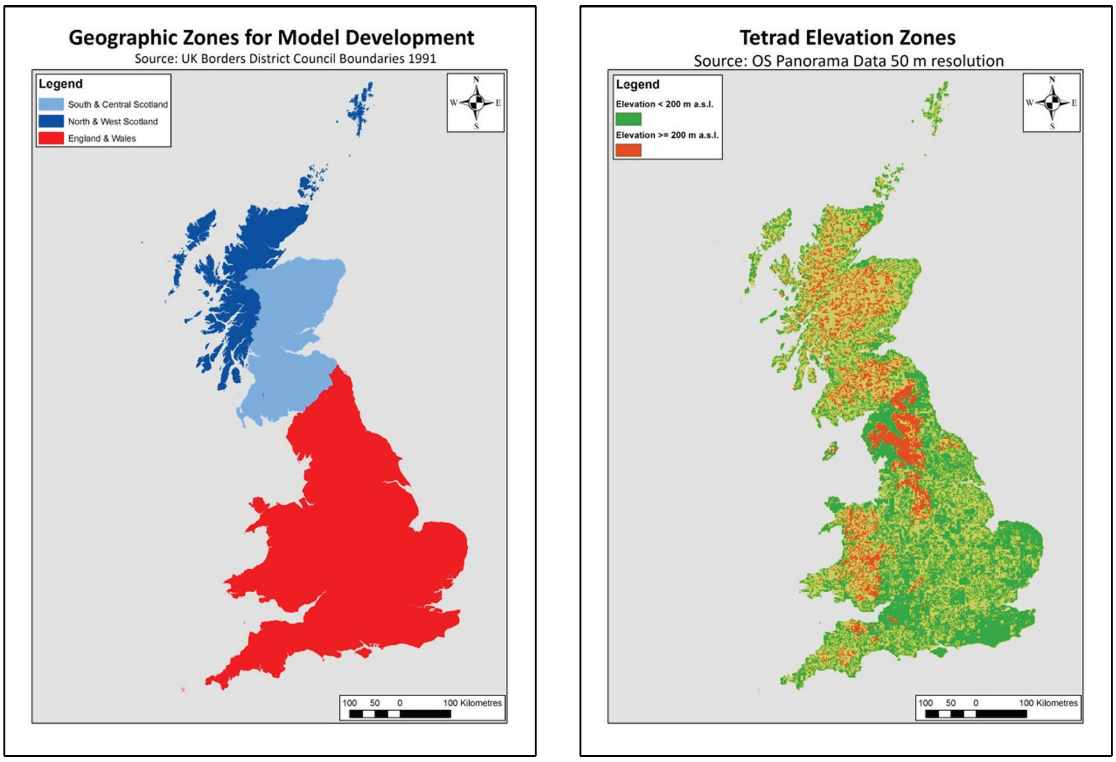

CLC90 Land Cover Data
LCS88 Land Cover Data
Data
• CLC90
• AgC88
• Elevation
• LTA Rain
• LA Temp
Averages
Calculate average comparison scores for CLC90 & AgC88
Averages
Calculate average comparison scores for CLC90 & AgC88
Data
• LCS88
• AgC88
• Elevation
• LTA Rain
• LA Temp
Combined data sets ready for Logistic Regression Analysis
Open in PASW
Correlation Analysis
Undertake correlation analysis of all variables to establish relationships & identify any collinearity
Logistic Regression Analysis
Undertake logistic regression analysis to determine an array of different models, with PRE, PGR & ZRE values saved to the original data set for each model
Goodness of Fit Assessment
Undertake assessment of various measures of 'goodness of fit' including -2LL, Chi-Square, Area under the ROC curve and AIC
Open in MS Excel
Adjustment of Presence/Absence Threshold
Adjust the cut-off value for the presence of twite to produce same number of tetrads with twite as the original BTO88 data
Open in ArcGIS
Map Results
Export results to ArcGIS to present results in map form and compare predicted with actual distribution

Figure 17 – Illustration of Logistic Regression Process

A progression of increasingly complex approaches was made to developing a predictive model for twite distribution. Initially a relatively simple model was derived based on the BTO88 twite tetrad data (32,784 tetrads with twite present in 1,044), the CLC90 land cover data and the AgCensus88 land use data. This was then elaborated to include elevation, and then climate. As a result of apparent variations in the accuracy of the models across Britain (see sections 3.3-3.5) the raw data was then divided into three distinct geographic zones: north and west Scotland (NWSco), south and central Scotland (SCSco) and England & Wales (EW) (Figure 18). The whole logistic regression model process was then repeated for each zone and the results amalgamated prior to exporting them to ArcGIS. The geographic zones were selected using boundaries from UKBORDERS District Boundaries and hence were convenient, but not necessarily related to actual processes and relationships between the variables. This could potentially result in Modifiable Arial Unit Problems (MAUP) (De Mers 2003, Heywood et al 2006, Longley et al 2005) where the results are more dependent on the boundary selection. So, finally, another zonation was carried out, this time using elevation as the critical factor (Figure 18). A quick assessment of the distribution of twite against elevation was undertaken, which suggested a cut-off of around 200 m might be appropriate. This reflected apparent differences in the relationship between elevation and twite presence in the coastal north & west as opposed to the more upland south & central Scotland and England & Wales. Two elevation zones were created, areas below 200 m and areas at or above 200 m. The whole logistic regression procedure was then repeated once more and the output combined for Britain as a whole.

One concern with this approach was the assumption that the land cover categories were reliable measures of habitat for twite. It was considered that the importance of heather moorland to the birds was perhaps not well represented by the generic moorland category within the Corine land cover data. Therefore, a separate model development procedure, mirroring the above, was undertaken separately for Scotland using the more detailed LCS88 land cover data in place of the CLC90 data. The LCS88 data consists of a very different classification of land cover and includes a specific category called *heather*. LCS88 also includes other potentially relevant categories such as *arable*, *improved grassland* and *bracken*.



Geographic Zones for Model Development

Source: UK Borders District Council Boundaries 1991

Tetrad Elevation Zones

Source: OS Panorama Data 50 m resolution

Figure 18– Geographic & Elevation Zones used for model Development

2.2.4 Model Development – Habitat Proximity

In addition to the logistic regression analysis described in the preceding section an attempt was made to introduce a measure of habitat proximity and interaction into the models. This was done using the Habitat Proximity variables described in section 2.2.2 in conjunction with the output from the logistic regression models. In essence the Habitat proximity variable was used as a form of weighting to adjust the model output in favour of areas where moorland was within 1.6 km of farmland as defined in section 2.2.2. This calculation was undertaken in ArcGIS using the ***Spatial Analyst – Raster Calculator*** tool to multiply the predicted probability by various factors (e.g. 0.85) and the Habitat Proximity by the remaining factor (e.g. 0.15) and adding the two products together (see equation 2.4 below). Thus the higher the Habitat Proximity value the greater the weighting applied to the model output

$$PRE_{Habitat} = (PRE_{LogReg} \text{ x } 0.85) + (\textit{Habitat Proximity} \text{ x } 0.15) \qquad (2.4)$$

where, PRE_{Hab} is the adjusted probability of twite weighted by the proximity of key habitats to each other, PRE_{LogReg} is the predicted probability of twite from the logistic regression model, and *Habitat Proximity* is the measure of moorland within 1.6 km of enclosed farmland.

2.2.5 Model Testing – Receiver Operating Characteristic Values (ROC)

The product of the model development was a number of different models, based on a combination of a range of parameters, logistic regression and proximity of key habitats. A critical requirement was to identify objectively, and preferably quantitatively, which of these models produced the best results compared with the actual twite data. It was decided to use the Receiver Operating Characteristic value (ROC), and specifically the area under the ROC curve, as a measure of performance. This choice was made on the basis that it was a relatively simple procedure to undertake, the analysis could be carried out on all the models, not just on the logistic regression components as was the case for the other standard statistical measures of 'Goodness of Fit', and it produces a standardised quantifiable value that allows the models to be directly compared.

The ROC curve is defined by Park et al (2004) as a plot of test *sensitivity* as the y coordinate versus its *1-specificity*, or false positive rate (FPR), as the x coordinate (see Figure 19), and is considered an effective method of evaluating the performance of diagnostic tests (Chan 2004, Park 2004) commonly used in medical research (Ganfyd 2010). *Sensitivity* is defined as the number of true positive decisions divided by the number of actually positive cases and *specificity* is defined as the number of true negative decisions divided by the number of actually negative cases (Park et al 2004). The area under the curve gives a quantitative indication of how good the test is. The ideal curve has an area of 1, the worst case scenario is 0.5 which equates with a performance no better than random chance (Ganfyd 2010).

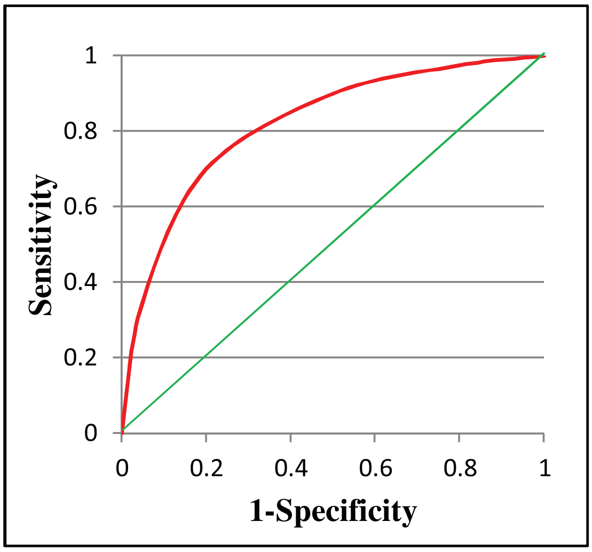


Figure 19 – Illustrative Example of a ROC Curve

The red line in Figure 19 represents the ROC curve, while the green line represents a value of 0.5, i.e. random chance. The larger the gap between the two lines the better the performance of the model. A ROC curve analysis was carried out using PASW software on the all model predictions, both predicted probabilities (*PRE*) and adjusted predicted groupings (*PGR*). The results of this analysis were used to determine which models performed best.

2.2.6 Model Testing – Independent Data

It is generally considered wise to test models against independent data, i.e. data that is similar to that used in development of the model, but is either from an independent source or is a sample from the same source, but which was excluded and not used in the development of the model. In this case the amount of presence/absence data for twite was severely limited but the 1999 National Survey data did provide a small data set that was suitable for this purpose. The best models were run using CLC2000 and AgCensus2000 data along with the same elevation and climate data, and the results were compared with the twite data from 1999. Comparisons were undertaken both visually in map form and by calculating the Area Under the ROC Curve statistic. Unfortunately there was no independent data available to specifically test the LCS88 models so this was not carried out.