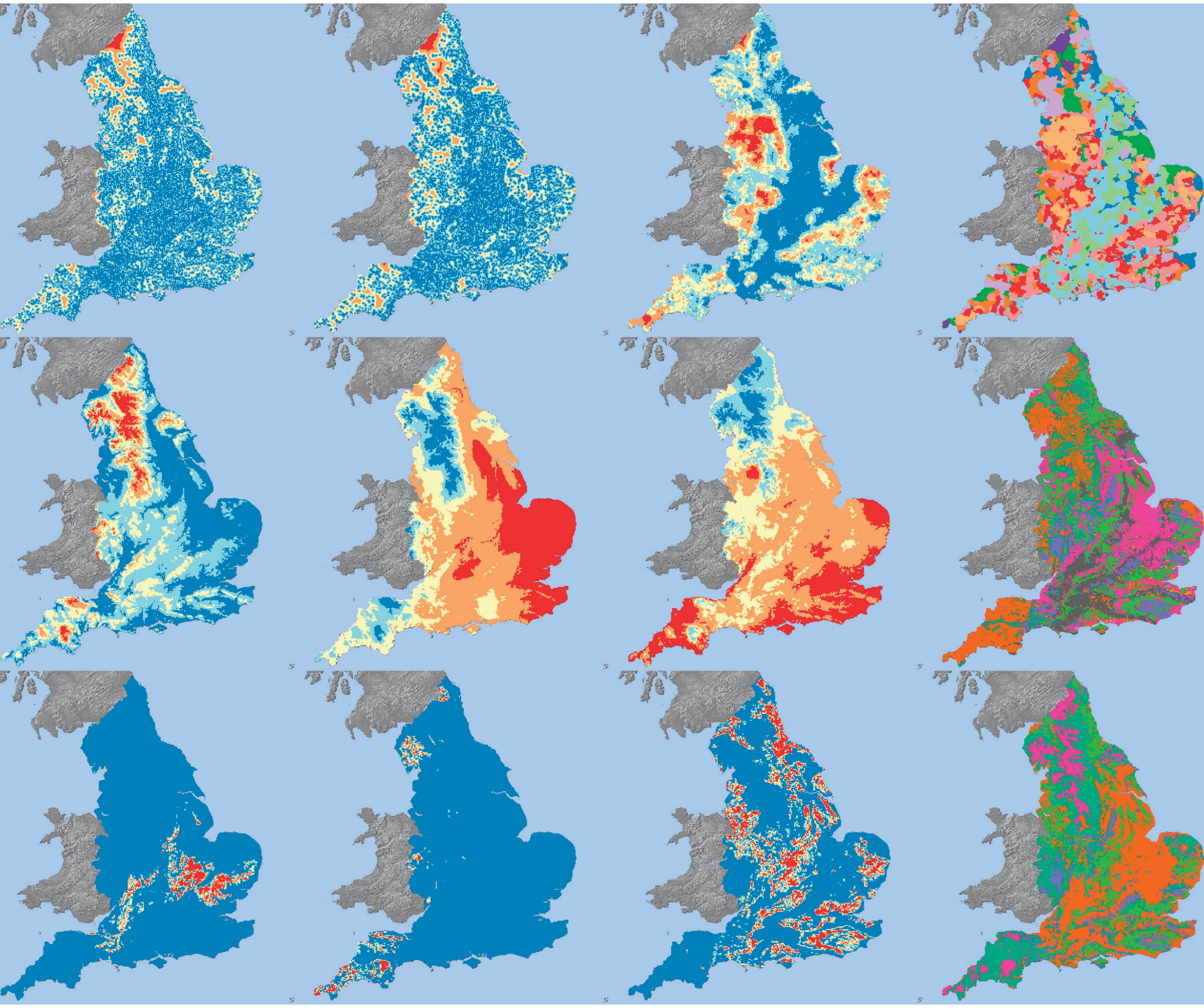
RESEARCH REPORT SERIES no. 72-2014

RURAL SETTLEMENT IN ENGLAND
ANALYSING ENVIRONMENTAL FACTORS AND REGIONAL VARIATION IN HISTORIC RURAL SETTLEMENT ORGANISATION USING REGRESSION AND CLUSTERING TECHNIQUES

Andrew Lowerre



INTERVENTION
AND ANALYSIS

ENGLISH HERITAGE

Rural Settlement in England Analysing Environmental Factors and Regional Variation in Historic Rural Settlement Organisation Using Regression and Clustering Techniques

Andrew Lowerre

ISSN 2046-9799 (Print)
ISSN 2046-9802 (Online)

The Research Report Series incorporates reports by the expert teams within the Investigation & Analysis Division of the Heritage Protection Department of English Heritage, alongside contributions from other parts of the organisation. It replaces the former Centre for Archaeology Reports Series, the Archaeological Investigation Report Series, the Architectural Investigation Report Series, and the Research Department Report Series.

Many of the Research Reports are of an interim nature and serve to make available the results of specialist investigations in advance of full publication. They are not usually subject to external refereeing, and their conclusions may sometimes have to be modified in the light of information not available at the time of the investigation. Where no final project report is available, readers must consult the author before citing these reports in any publication. Opinions expressed in Research Reports are those of the author(s) and are not necessarily those of English Heritage.

Requests for further hard copies, after the initial print run, can be made by emailing:
Res.reports@english-heritage.org.uk
or by writing to:
English Heritage, Fort Cumberland, Fort Cumberland Road, Eastney, Portsmouth PO4 9LD
Please note that a charge will be made to cover printing and postage.

SUMMARY
This report shows how it is possible to collate Geographic Information Systems (GIS) data for historic settlement nucleation and dispersion with a range of data on environmental variables in order to investigate the relationships between them.

Ordinary Least Squares (OLS) regression model specification, selection and validation procedures, followed by further analysis using spatial regression methods, identified environmental variables that appear to have had the most significant influence on settlement organisation. The use of OLS and spatial regression and the innovative Relative Area Overlap (RAO) technique has enabled investigation of how relationships between key environmental variables and historic settlement organisation varied across England. Overall, the regression analyses indicate that far more of the variation in the measures of settlement organisation is not explained by the environmental variables than is explained by them. The results of the RAO analysis echo this conclusion.

Using unsupervised classification, it has been possible to develop new, national-scale characterisations of historic settlement organisation and of key environmental variables. These new classifications of historic settlement organisation often broadly align with Brian Roberts and Stuart Wrathmell's delineations of provinces, sub-provinces and local regions, but the cluster outlines and Roberts and Wrathmell's boundaries diverge more often than they agree.

CONTRIBUTORS
All data manipulation, analysis, mapping and report-writing for this project was undertaken by Andrew Lowerre.

ACKNOWLEDGEMENTS
The research presented here would not have been possible without the support and considerable patience of Jeremy Lake, Michael Russell and Brian Kerr.

ARCHIVE LOCATION
The digital archive for this project is currently held at Fort Cumberland. The settlement data and related environmental datasets will, where possible due to copyright and other intellectual property rights restrictions, be deposited with the Archaeology Data Service (ADS).

DATE OF STUDY
November 2012 – October 2014

CONTACT DETAILS
English Heritage, Fort Cumberland, Fort Cumberland Road, Eastney, Portsmouth, PO4 9LD
Andrew Lowerre Tel: 02392 856765. Email: andrew.lowerre@english-heritage.org.uk

CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

The nature and causes of regional variation in rural settlement organisation in England from the early medieval period onwards have long been subjects of study. The basic contrast is between areas of dispersed settlement (where houses, farms, churches and so on are spread widely across the landscape) and areas of nucleated settlement (where such structures stand in compact groups), recognising that the gradations between the two forms of organisation are often quite subtle. A variety of factors which can be simply, if rather crudely, labelled 'environmental' and 'cultural' have been offered to explain the variations, but no one factor can be identified as an obvious 'prime mover' in all times and in all places. Williamson (2010) gives a useful summary of the scholarly debates developed over the last century of study.

One of the key points of reference for understanding the development of rural settlement in England and the historic character of the landscape is Roberts and Wrathmell's *An Atlas of Rural Settlement in England* (2000). Their aim was to portray complex patterns of settlement organisation at a national scale. Working from nineteenth-century Ordnance Survey 'Old Series' 1:63,360 (one inch to one mile) scale maps and using a method involving, as they put it, 'little science but much logic' (Roberts and Wrathmell 2000, 13), they delineated a comprehensive, hierarchical set of settlement provinces, sub-provinces and local regions. Their 'Central Province,' running from south-west to north-east across the central part of the country, is dominated by nucleated settlement. The 'South-Eastern' and 'Northern and Western' provinces lying either side of the Central Province are largely characterised by more dispersed settlement. Sub-provinces and local regions exhibit more nuanced, locally-focused variations within the broad trends in each province.

A recent study by Lambourne (2010) examines patterning in the historic landscape of south-western England, investigating both environmental and cultural factors, and offering a sustained (though not wholly negative) critique of Roberts and Wrathmell's methods and the delineation of their Central Province. Investigation either side of the south-western boundary between the Central and Northern and Western provinces has emphasised the significance of cultural factors in shaping settlement patterns over 'the long eighth century' (Rippon 2010; Rippon *et al* 2006). Jones (2010, 36-40) discusses the potential effects differences in topography and soils in Roberts and Wrathmell's three settlement provinces might have had on the ongoing and sometimes chaotic development of settlement organisation in England. The importance of environmental factors in influencing the development of nucleated settlements in the medieval period has been emphasised most strongly by Tom Williamson (2003; 2005; 2010; 2013). The project reported on here focuses on the environmental factors highlighted by Williamson

Previous work investigating the relationships between environmental variables and variation in historic settlement organisation has been regionally focused, rather impressionistic (particularly when done at a national scale), or both. Williamson's model of the influence of environmental factors (including soils, precipitation, topography, the availability of land suitable for creating meadow and so on) on patterns of medieval settlement organisation appears elegant and compelling, but it is not clear how well that model explains variation in settlement organisation across the whole of England. Williamson also advocates the use of digital mapping and GIS (Williamson 2013), but his own use of these technologies does not appear to extend beyond basic data management and cartography.

In this project I took an explicitly spatial statistical approach, operating at a national scale. The goal was to quantify the strength of the relationships between environmental factors and regional variation in historic settlement organisation and show how and where those relationships fluctuated across England. I also sought to identify at what scales variation in settlement organisation is most apparent and at what scales the relationships between environmental factors and settlement variation are clearest. The application of sophisticated statistical and spatial analytical techniques is only now possible because of the conversion of Roberts and Wrathmell's settlement nucleation and dispersion maps into geospatial data (Lowerre 2010; Lowerre *et al* 2011), along with the increasing availability of a range of GIS-ready environmental data.

My intent was not to advocate a completely 'environmentally determinist' approach to understanding variation in historic settlement organisation. It was, rather, to explore the strength (or weakness) of the relationships between environmental factors and settlement organisation and how those relationships vary across the country. 'Cultural' factors (eg, farming methods, forms of land tenure, inheritance practices, the effects of human activity on soils and the historical contingencies of earlier settlement to name only a few) obviously influenced the development of settlement organisation. Consideration of such factors was, however, firmly outside the scope of the project. It has, however, been possible to identify areas where environmental factors provide a poor explanation for settlement variation.

The GIS-based statistical and spatial analytical approaches deployed here have rarely been applied in historical settlement studies in Britain, so some justification for their use is perhaps necessary. It may be noted, for example, that Williamson does not present any maps directly overlaying soils, temperature or rainfall with those areas of England dominated by nucleated settlement and/or open fields in any of his works arguing the importance of environmental factors to variation in settlement organisation (Williamson 2003; 2005; 2010; 2013). Readers are left to flip back and forth between maps on different, often widely separated pages, making it difficult to judge how strong any patterns in or relationships between the datasets might be. The recent work of Williamson, Liddiard and Partida (2013) on Northamptonshire also makes extensive use of GIS for data management and cartographic purposes, but they do not appear to have employed any of the formal spatial analytical techniques now widely available in various GIS software packages. Similarly, Lambourne's (2010) study is based on a great deal of cartographic synthesis and evaluation but is entirely non-quantitative.

There is, however, a wide range of evidence indicating that human beings are not always successful in correctly identifying patterns or the genuine absence of pattern (Whitson and Galinsky 2008). MacEachren (1995, especially 435-58) discusses the extensive literature on the visual recognition and interpretation of patterns in maps. Klippel *et al* (2011) review more recent work and highlight the difficulties in interpreting observed spatial patterns with regard to randomness and statistical significance. The questions posed in this study cannot be answered without being able to recognise reliably genuine spatial patterns in the data and associations between different sets of data. Particularly when working with relatively high-resolution, multivariate data that are national in extent, there is a considerable risk that simply 'eyeballing' the data may suggest patterns or relationships that are not actually there (false positives) as well fail to recognise patterns and relationships that are genuinely present (false negatives) (see, inter alia, Muller 1975; 1976; 1979; Evans 1977; Brewer and Pickle 2002). Spatial statistical analysis makes it possible to summarise patterns and relationships between and within datasets, to examine how and where those relationships vary, and to test the likelihood that any patterns or relationships are the result of random chance (Gregory 2008).

The overall approach taken here has been heavily influenced by methods used in the fields of landscape ecology, biogeography and spatial epidemiology, examining distributions of plant or animal species (or different inter- and intra-species traits) or disease cases in relation to a range of environmental variables. In the same spirit as Bevan (2012), I sought to apply, and where necessary, develop robust methods for dealing with spatially extensive data while addressing a fundamental question about past human behaviour. More generally, the project aimed to address a fundamental methodological question posed by Bevan and Conolly (2009, 956): 'how, ultimately, do we identify and make sense of the heterogeneous and often inter-dependent behaviours and processes responsible for apparent spatial patterns?'.

PROJECT AIMS AND OBJECTIVES AND BUSINESS CASE

The project had three aims, each with a set of supporting objectives, posed as research questions.

Aim 1: To investigate the inter-relationships of environmental factors and historic settlement organisation, and how they are expressed as regional and local variations

- Objective 1.1: Can GIS data for historic settlement nucleation and dispersion be collated with a range of data on environmental variables?
- Objective 1.2: Which environmental variables (if any) appear to have had the most significant influence on regional variation in historic settlement organisation?
- Objective 1.3: How did the relationships between the key environmental variables and historic settlement organisation vary across England?

Aim 2: Develop a new, national-scale characterisation of historic settlement organisation as it relates to the physical environment

- Objective 2.1: Is it possible to create a national classification of the landscape based on variations in historic settlement organisation and key environmental variables?
- Objective 2.2: Assuming Objective 2.1 can be met, how does the new classification compare with Roberts and Wrathmell's delineation of settlement provinces, sub-provinces and local regions?

Aim 3: Share the results of the project through publications and presentations and archive and disseminate the analytical data produced

- Objective 3.1: Publish an article on the results of the project in a peer-reviewed journal and in an English Heritage Research Report Series report
- Objective 3.2: Publish notes on the project on a project website and in *Research News* (or similar)
- Objective 3.3: Present the results of the project at one or more relevant professional/academic conferences and/or workshops
- Objective 3.4: Archive and disseminate the analytical data via the English Heritage Archive and the project website, as well as through the Archaeology Data Service (ADS)

The project primarily contributes to the National Heritage Protection Plan's Protection Result 4F1.1: Strategic guidance and assessment tools for protection through managing change to rural buildings and their settings (English Heritage 2011a; English Heritage 2011b). Settlement, its organisation and environmental characteristics help form and inform perceptions of local distinctiveness and a 'sense of place'. Better understanding of how environmental factors contributed to regional variation in historic settlement organisation can improve understanding of the significance of variation in rural settlement and their historic character, as well as the overall setting of many types of rural buildings.

The project also uses novel approaches to the study of historic rural settlement organisation and so represents a significant methodological advancement. The project has capitalised on investment already made by English Heritage, both in the original *Atlas* and in the GIS dataset produced from it. It has begun to realise some of the enormous research potential embodied in the Atlas of Rural Settlement in England GIS dataset.

DATA SOURCES AND DATA PREPARATION

Settlement Data

Rather than use the boundaries of the settlement provinces, sub-provinces and local regions delineated by Roberts and Wrathmell as the basis for analysis, I used the building blocks from which they derived their boundaries: the locations of nucleated settlements of various sizes and the quantifications of settlement dispersion, their dispersion scores and hamlet counts.

As noted above, the mid-nineteenth-century settlement nucleation and dispersion data are those presented in Lowerre et al (2011). The nucleation and dispersion data are represented as points, so a variety of methods were used to calculate values for the whole of England based on the locations and values of the known features (cf. Lowerre 2010, 33-40). The basic approach is to divide the whole of England into a grid or lattice of regularly-spaced cells and, working from the known data, calculate a value for each cell based on the values of the points nearest to it. I used a grid cell size of 2 x 2km, mirroring the size of the sample areas Roberts and Wrathmell used when quantifying settlement dispersion (Roberts and Wrathmell 2000, 12-13).

Recognising that some scholars have raised questions about the reproducibility of Roberts and Wrathmell's results (Dyer 2003; Hinton 2005), I have assumed that the data on nucleations are (for the purposes of this study, at least) complete and correct, ie, they did not leave out any nucleations and all the nucleations they mapped really existed. There are eight nucleation and six dispersion score/hamlet count sample points which lie outside the current boundary of England – just over the Welsh border – in the Atlas GIS dataset. These points were included when generating surfaces from the settlement data, but the full range of environmental data for these locations was not available (see the discussion of environmental data below). The grid cells in which these nucleation and dispersion score/hamlet count points lie are not present in the final dataset used for analysis.

Nucleations

I calculated the simple, Euclidean distance from the nucleation points to the centre of each 2 x 2km grid cell to provide a measure of the concentration of nucleated settlement or 'nucleatedness'. The distance was calculated for all nucleations, regardless of the size class (A-E) assigned by Roberts and Wrathmell. I also calculated distances for nucleations of each class on its own, as well as for nucleations in categories B, C or D taken together.

Calculating the simple distance to nucleations is preferable to calculating the density of nucleations as a measure of the concentration of nucleations because of the problems caused by edge effects. The difference in the number of points along the coast and the borders on the one hand and the number of points in central England on the other is such that the density along the 'edges' is profoundly suspect. Furthermore, computing 'density of nucleation' as a continuous surface is arguably illogical if one assumes the nucleations data are complete and each nucleation should be treated as an indivisible unit.

Dispersion Scores

The dispersion score points in the Atlas GIS dataset represent the locations of the 2 × 2km sample areas where Roberts and Wrathmell counted the houses, farmsteads, cottages and so on depicted on the Old Series Ordnance Survey maps they used as their source. Obtaining values for non-sampled locations from the known samples required the estimation or interpolation of continuous surfaces for visualisation and analysis. A wide range of different interpolation methods is available (Lloyd 2009, 129–54).

Interpolating the dispersion scores as recorded by Roberts and Wrathmell was complicated by the numeric scale they used. Roberts and Wrathmell took counts from their sample areas, but then constrained the actual counts to fit the Fibonacci sequence, where each value in the sequence is the sum of the two preceding values (0, 1, 2, 3, 5, 8, 13, 21, 34 and so on) (Roberts and Wrathmell 2000, 12-13). The number scale in the data as presented in the published *Atlas* and in the GIS dataset derived from it is, therefore, neither truly interval nor truly categorical. Most spatial interpolation methods assume the input data are recorded on at least an interval scale. Applying such interpolation methods to the dispersion score data in their 'raw' form would produce unreliable results.

I addressed this issue by adding a random value to each observation with a score of 3 or greater and then interpolating from the 'randomised' values. The random values added to each point were limited to match the 'missing' values left out in the Fibonacci sequence. For example, if the value for a given point was 5, a value of 0, 1 or 2 was randomly added; if the recorded value was 8, a value of 0, 1, 2, 3 or 4 was randomly added. I only added values because of the 'if in doubt round down' approach taken by Roberts and Wrathmell when they originally compiled the counts (Roberts and Wrathmell 2000, 12). I repeated this process ten times to produce ten sets of randomised dispersion scores where the scores encompassed the full range of integers from 0 to 55. I used these ten sets as training data for exploring which method worked best for interpolating the dispersion scores.

I applied a variety of different interpolation methods using varying parameters to the ten training sets, including Inverse Distance Weighting, Local Polynomial Interpolation, Kernel Smoothing (without barriers), Thin Plate Splines, Thin Plate Splines with Tension, Completely Regularised Splines, and Multiquadric radial basis functions. I used the Geostatistical Analyst extension in ArcGIS 10.0 to create all the interpolated surfaces. Detailed descriptions of how each interpolation method works can be found in Lam (1983), Burrough and McDonnell (1998, 98-121), Lloyd (2010, 145-68) and Chang (2012, 314-324). The implementations of the interpolation methods in the ArcGIS 10 Geostatistical Analyst are summarised in Esri's online Help files (Esri 2011a).

I employed cross-validation to assess how successfully each interpolated model predicted values at unsampled locations, as well as to compare the results from different models. Cross-validation works by removing one observation from the sample set and predicting the data value at that sample location from the remaining samples. The process is repeated, one sample at a time. The actual value at each sample point is subtracted from the predicted value to determine the prediction error, and a series of statistics are calculated to summarise the errors (Bailey and Gatrell 1995, 191; Chang 2012, 333-35). I compared the mean error, root mean squared error, and, where applicable, the mean prediction standard error, the mean standardised prediction error and the root mean squared standardised prediction error for each interpolated model. The Thin Plate Splines with Tension method, using a kernel parameter of 0.007634, a minimum of 4 and a maximum of 32 neighbours and a search radius of 24km, produced the best results.

I then created 100 new randomised sets of dispersion scores and produced interpolated surfaces from each iteration, using the Thin Plate Splines with Tension method just described. I then converted the 100 interpolated surfaces to a 2 x 2km grid and averaged those surfaces to get an approximated version of a surface interpolated from 'true' counts. The interpolation process produced continuous numeric values rather than integers, but the dispersion scores as originally captured by Roberts and Wrathmell are effectively counts, that is, whole numbers. To reflect the nature of the original scores, I rounded the value for each grid cell to the next lowest integer value, again following Roberts and Wrathmell's 'when in doubt, round down' approach.

Hamlet Counts

Interpolating the hamlet count data was, in comparison to the dispersion score data, much simpler. The hamlet counts are true counts, so it was possible to work directly from the data in the Atlas GIS dataset. I applied the Inverse Distance Weighting, Local Polynomial Interpolation, and Completely Regularised Splines methods, using varying parameters. Again, I used cross-validation statistics to assess and compare the results from each method. The Inverse Distance Weighted method produced the best result, using the simple inverse distance, a minimum of 10 and a maximum of 15 neighbours within a search radius of 32km. As with the dispersion scores, I converted the interpolated surface to a 2 x 2km^2 grid and rounded the continuous numeric values to the next lowest integer.

Combined Settlement Scores

To allow the visualisation and analysis of all three settlement variables simultaneously, I derived a series of Combined Settlement Scores (CSS), based on the interpolated values for distance to nucleations, dispersion scores and hamlet counts. To enable the use of various quantitative analytical methods, I calculated the CSS values as continuous numeric data, rather than as categorical data, as in earlier work (Lowerre 2010, 36, 38-40).

The values for the distance to nucleations, dispersion score and hamlet count variables differ considerably in their magnitude and scale. Distances to nucleations range from 17.9m to 26,943.2m, dispersion scores from 0 to 26, and hamlet counts from 0 to 8. To give each of the variables equal weight when calculating CSS values, I standardised the variables to a scale of 0.0 to 1.0 by subtracting the minimum from each value and then dividing by the value range (Milligan and Cooper 1988, 185).

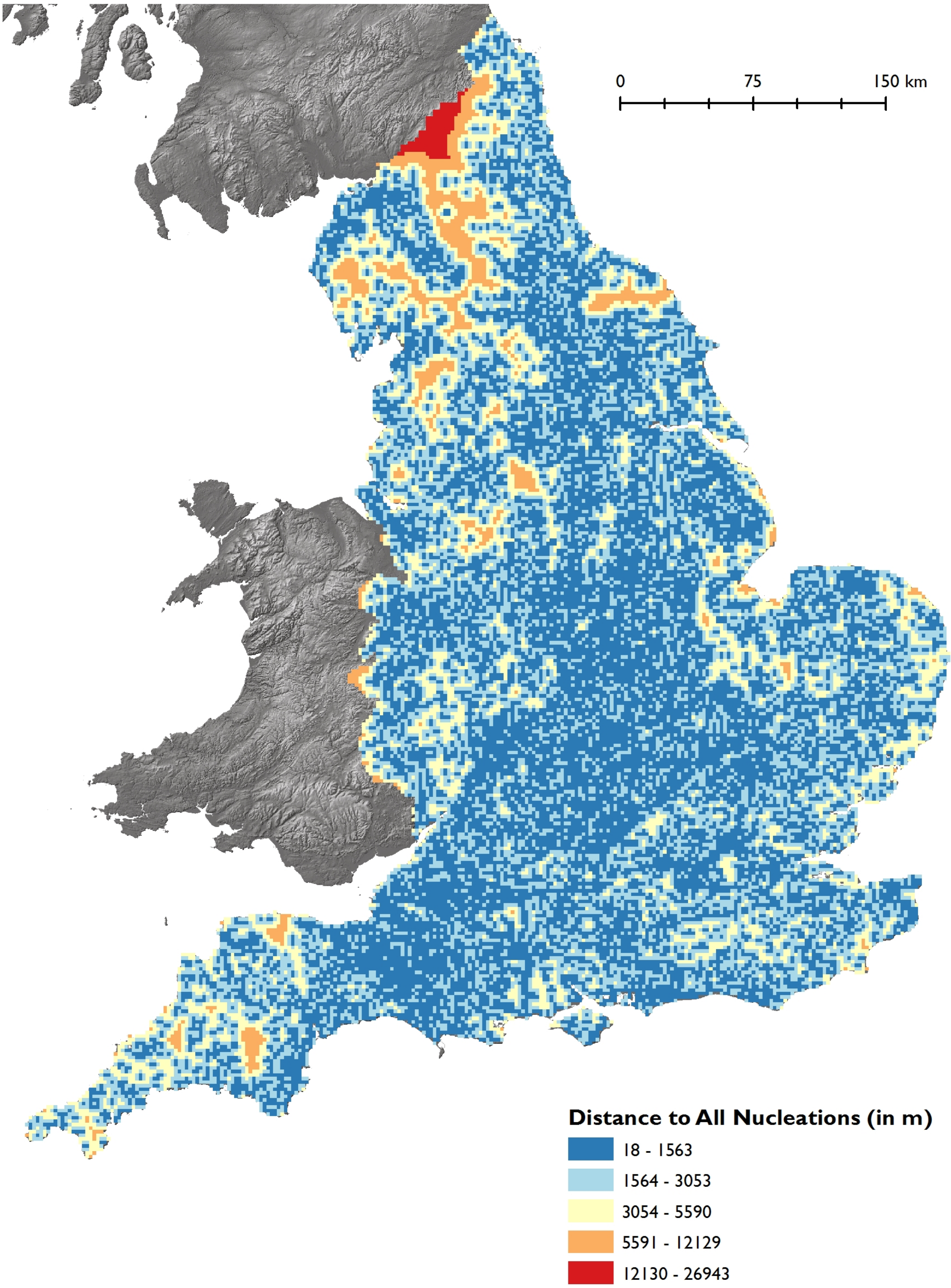
I computed the CSS value as the average of the standardised nucleation, dispersion and hamlet count scores for each grid cell. I calculated eight different versions of CSS, depending on whether the CSS value was intended to emphasise nucleation or dispersion, whether the distance to all nucleations was used or only the distance to B, C and D category nucleations, and whether hamlet counts are seen as contributing to greater nucleation or greater dispersion. Table 1 illustrates the different versions of CSS and how I calculated them. Example values for standardised nucleation, dispersion score and hamlet count variables are given in brackets after each column heading, and the resulting CSS values are given in the last column on the right. It should be noted that the respective N and D variants of CSS are the inverse of each other.

Table 1: Combined Settlement Score (CSS) versions and example values

	Nucleation (a = 0.158; b = 0.236)	Dispersion Score (0.379)	Hamlet Count (0.125)	CSS Value
Na1	Inverted	Inverted	Normal	0.529
Nb1	Inverted	Inverted	Normal	0.503
Na2	Inverted	Inverted	Inverted	0.779
Nb2	Inverted	Inverted	Inverted	0.753
Da1	Normal	Normal	Inverted	0.471
Db1	Normal	Normal	Inverted	0.497
Da2	Normal	Normal	Normal	0.221
Db2	Normal	Normal	Normal	0.247

N = emphasises nucleation (ie, locations closer to nucleations have a higher score)
D = emphasises dispersion (ie, locations further away from nucleations have a higher score)
a = uses distance to all nucleations
b = uses distance to B, C and D nucleations
1 = Hamlet Count contributes toward nucleation (ie, locations with a higher Hamlet Count are considered to be more nucleated)
2 = Hamlet Count contributes toward dispersion (ie, locations with a higher Hamlet Count are considered to be more dispersed)

The maps on the next pages illustrate the distribution of values for distance to all nucleations, distance to nucleations in categories B–D, and CSS variants Na2 and Nb2. In all four maps, I grouped the values into five classes using Jenks's Natural Breaks method (Jenks and Caspall 1971; Jenks 1977) as implemented in ArcGIS (Esri 2012).



Contains Ordnance Survey data © Crown copyright and database right 2014

Figure 1: Map of distance in metres to all nucleations (categories A–E)

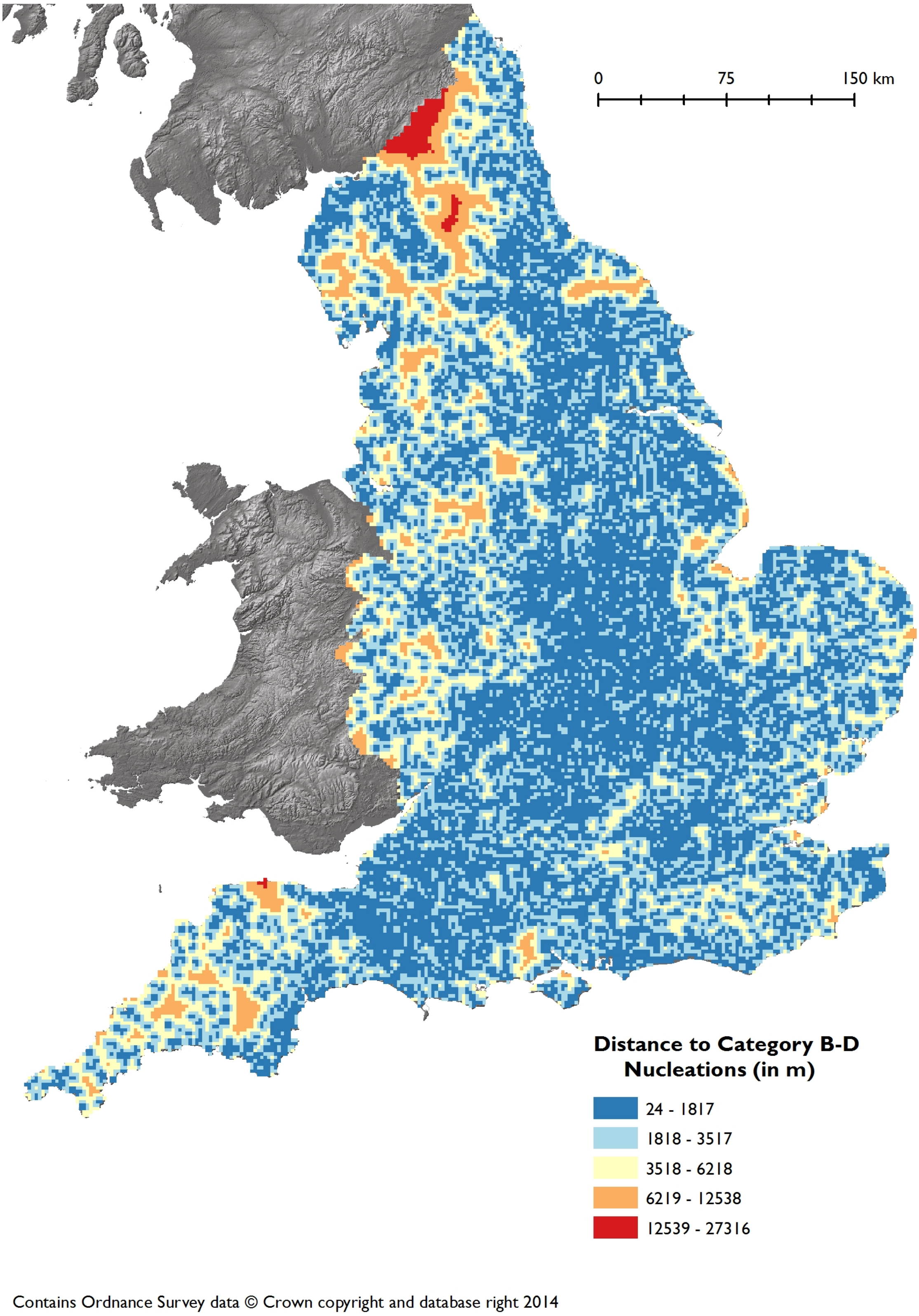
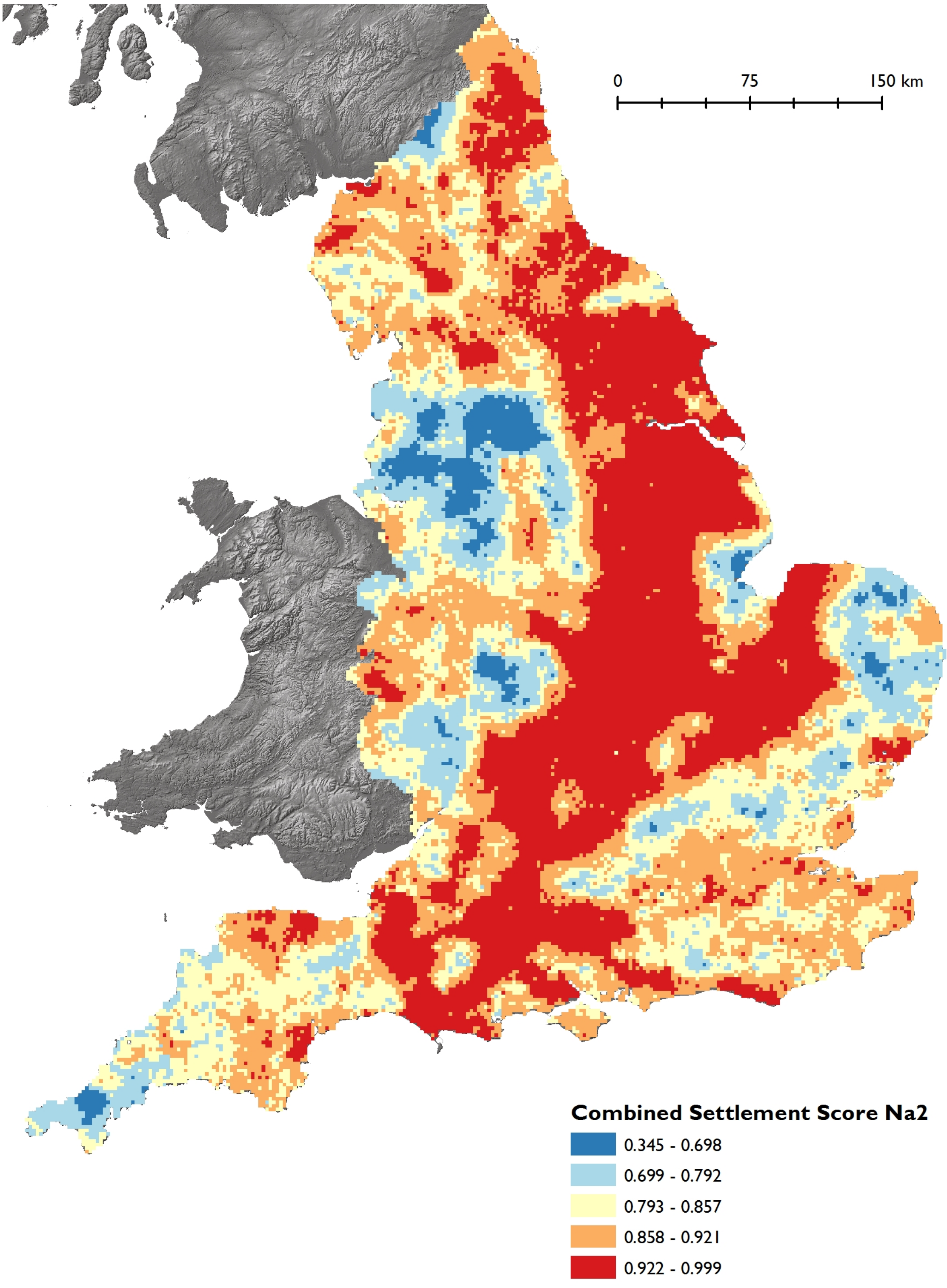
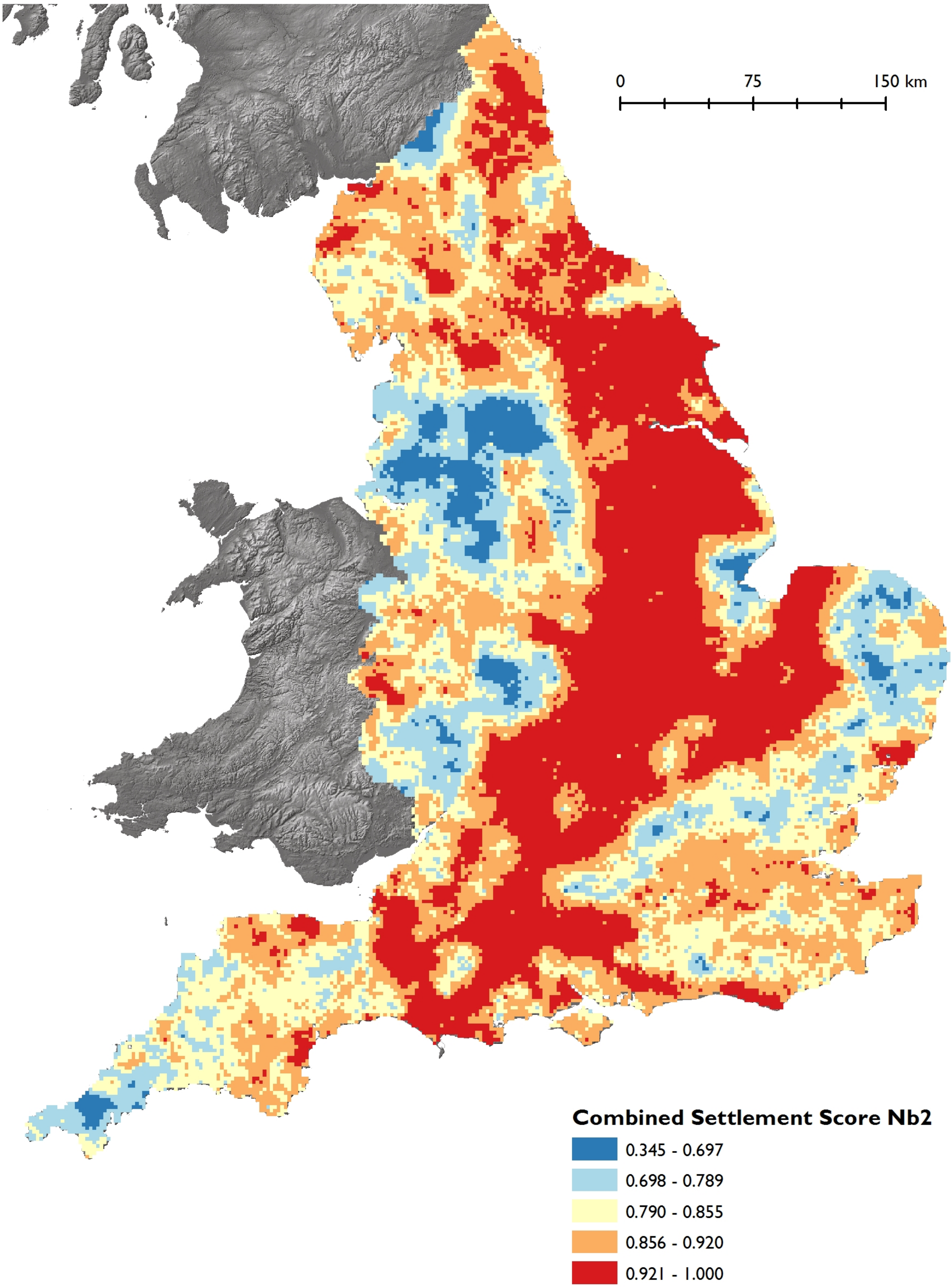


Figure 2: Map of distance in metres to category B–D nucleations



Contains Ordnance Survey data © Crown copyright and database right 2014

Figure 3: Map of Combined Settlement Score Na2



Contains Ordnance Survey data © Crown copyright and database right 2014

Figure 4: Map of Combined Settlement Score Nb2

Environmental Data

I used data on soils, elevation, surface roughness, precipitation, temperature and solar radiation as environmental variables in the analysis.

Soils

Data on soils in England were taken from the National Soils Research Institute's *NATMAP Soilscapes* dataset (National Soils Research Institute (NSRI) 2001; 2013). Soilscapes is a vector polygon dataset at a nominal scale of 1:250 000, representing in simple terms likely soil conditions across the country. The dataset divides soils into one of twenty-seven broad types. I imposed a 2 x 2km vector grid over the data and calculated the area of each grid cell occupied by each soil type. I also calculated for each grid cell the proportion of the total area of the cell covered by the soils data occupied by each soil type. I excluded from the dataset Soilscape types 0 (sea), 28 (water) and 29 (unclassified).

In his discussions of the influence of soils on past people's decisions regarding settlement organisation, Williamson focuses on a number of soil associations (eg, Williamson 2003, 63-65, 124-25, 142-47; Williamson 2013, 46-51). Soil associations provide a more detailed and spatially refined classification of soils than that presented in the Soilscapes dataset. NSRI license GIS-ready data on soil associations, but these data were not available to this project because of cost. The Soilscapes data are, however, derived directly from the soils associations, and a list of the associations comprising each Soilscapes type is presented in Appendix 2.

Elevation and Surface Roughness

I used Ordnance Survey's Landform PANORAMA 50m-resolution data (Ordnance Survey 2010) to construct a digital elevation model for the whole of England. The digital elevation model (DEM) records the elevation above sea level for each 50 x 50m grid cell.

To explore the possible effects of relative variation in local topography (in addition to simple elevation above sea level), I derived several measures of topographic variability or surface roughness from the elevation data: the Topographic Roughness Index (TRI) (Riley *et al* 1999), surface ratio (Jenness 2004), the Vector Ruggedness Measure (VRM) (Sappington *et al* 2007), relief, and standard deviations of elevation and slope (Grohmann *et al* 2011). I calculated the VRM, relief and standard deviations of elevation and slope measures each using three different sizes of moving window: 3x3, 11x11 and 21x21.

Precipitation

Data on rainfall were obtained from the WorldClim website (http://www.worldclim.org/) (Hijmans *et al* 2005). The WorldClim data give monthly long-term averages for precipitation (in mm) as well as overall annual averages, based on the period 1950-2000. The data are provided as a grid covering global land areas (excepting Antarctica), with a spatial resolution of 30 arc-seconds (approximately 0.93 × 0.93km at the equator). I also used data described by Perry and Hollis (2005a; 2005b) on the 1961–1990 long-term averages for monthly days of rain ≥ 1.0mm and 10.0mm in the United Kingdom, supplied by the Met Office as gridded data at a spatial resolution of 1 × 1km. The goal was to explore whether the frequency of episodes of heavy precipitation at particular times of year (rather than monthly averages for total precipitation) affected patterns of settlement organisation. In addition to the monthly values, I calculated moving two- and three-month averages for monthly days of rain and sums and averages for precipitation. The intent was to investigate whether trends over periods longer than a single month might have influenced settlement organisation.

Temperature and Insolation

Data on temperature were also obtained from the WorldClim website (http://www.worldclim.org/). Again, the WorldClim data give monthly and annual long-term averages, based on the years 1950-2000. For clarity, I recalculated the temperature data to be decimal degrees Celsius, as opposed to degrees Celsius * 10 as in the raw data.

I sought to mitigate the potential unsuitability of modern temperature data, in particular the 'urban heat island' effect (Oke 1982; Gallo *et al* 1996; Peterson and Owen 2005) through the use of measures of insolation rather than recorded air temperature. Insolation is the amount of solar radiation received by the landscape (Dubayah and Rich 1995; Fu and Rich 2002; McCune and Keon 2002). Using the ArcGIS Area Solar Radiation tool (Esri 2011b; Esri 2011d; Esri 2011e), I derived two measures of insolation from the elevation data discussed above: the amount of direct incoming solar radiation (in watt hours per m^2) and the duration of direct incoming solar radiation (in hours). Amounts of incoming solar radiation are heavily dependent on the latitude of the region analysed—all other things being equal, the amount and duration of solar radiation the landscape somewhere in Cornwall receives will be different to the amount and duration the landscape around Berwick-on-Tweed receives. To account for this effect, I calculated the insolation across England in a series of overlapping bands running from west to east. I then merged the results, averaging the values in the areas of overlap, to create insolation surfaces for the whole country. I calculated annual and monthly values for amounts and duration of solar radiation.

As with the precipitation and days of rain data, I also calculated moving two- and three-month averages for temperature and insolation. Again, the aim was to explore the possible influence on settlement organisation of trends over periods longer than a single month.

Discussion

The use of modern climate data to model nineteenth-century and earlier environmental conditions is less than ideal, but effectively unavoidable. Addressing the questions posed here required data at spatial resolutions considerably higher than those available in current, applicable palaeoclimate reconstructions (eg, Brewer *et al* 2007; Büntgen *et al* 2011; Bartlein *et al* 2011; Ljungqvist *et al* 2012). The development of appropriate, high-resolution palaeoclimate models was far outside the scope of the project. Other recent research relating climatic variables to archaeological evidence has made use of modern climate data, and like, for example, Conolly et al (2012, 1002), I have assumed '… with caution that the relative differences between regions have remained stable, and thus … measures of differences between regions also reflect past relative differences'.

The datasets described above come in a variety of resolutions and geographic projections. Where necessary, the datasets were reprojected onto the Ordnance Survey's British National Grid coordinate system. Pre-processing of the data (eg, deriving measures of surface roughness or insolation from the elevation data) was done with the data in their original resolution. Once all reprojection and pre-processing was complete, all the environmental data layers were resampled down to a 2 × 2km resolution, calculating the mean value for each 2 × 2km grid cell, to match the interpolated settlement nucleation and dispersion data.

Given their diverse origins, it is not surprising that the extents of the different environmental datasets do not align perfectly. As a result, there are locations along the coastline and the Welsh and Scottish borders where data for all variables are not available. Any grid cell for which values for all the environmental variables were not available was excluded from the analysis. The final dataset contains a total of 32,959 2 × 2km grid cells.

All data management, pre-processing and reprojection was carried out using Esri's ArcGIS versions 10.0 and 10.1.

ANALYSIS

I used a number of different analytical techniques to address the questions posed under Aims 1 and 2: non-spatial Ordinary Least-Squares (OLS) regression, spatial lag and spatial error regression, clustering analysis using unsupervised classification, and polygon-based relative area overlap analysis of the clusters produced using unsupervised classification. I examined four settlement variables: distance to category A–E nucleations (hereafter DstNclAll); distance to category B, C and D nucleations (hereafter DstNclBCD); and the Na2 and Nb2 Combined Settlement Scores (hereafter CSSNa2 and CSSNb2, respectively), that is, those in which locations closer to nucleations have a higher score than locations further away, and locations with a higher Hamlet Count are considered more dispersed.

Non-Spatial Ordinary Least-Squares Regression

Method

To investigate the relationships between the data on settlement organisation and various combinations of environmental factors, I began by developing a series of OLS multiple regression models. The method seeks to analyse values for one variable (known as the dependant or response variable) based on the values of one or more other variables (known as independent or explanatory variables) (see Burt *et al* 2009, 172-187, 498-508; Lloyd 2009, 31-39). Here, the interpolated values for concentrations of nucleations and Combined Settlement Scores are the response variables and the various environmental factors are the explanatory variables. OLS regression is a non-spatial method: it does not take into consideration the location of the points at which the response and explanatory variables are measured. The goal was to develop a small set of appropriately specified models, that is, to include all the most meaningful explanatory variables and leave out all those that do not contribute materially to explaining variations in historic settlement organisation.

Statistical modelling is often performed as an iterative process: the investigator fits (or specifies) a model to the available data, and on the basis of the results produced from that model, drops some variables and adds others to produce a new model, which he or she then inspects and uses as a basis for producing further models with yet more (or fewer) other variables, and so on. The investigator selects a 'final' model or small group of models that best fit the available data and bases his or her inference – in both the statistical and common sense meanings of the word – on the final model or models. There are, however, various problems with this approach – known as 'data dredging,' 'data snooping' or 'data mining' (in a pejorative sense) – which are widely recognised but often ignored (Lo and MacKinlay 1990; Chatfield 1995; White 2000; Burnham and Anderson 2002, 37-41, 72-74). There is a serious risk that apparently significant results may simply be the product of chance. A model found after extensive data dredging may fit very well the data used to create it, but that model may perform poorly when applied to new data. Ignoring the model specification process and basing one's inferences only on the final model – in effect acting as if the final model had been known all along – renders those inferences highly suspect. Best practice also entails checking or validating the models, to examine how robust the models' explanatory and/or predictive abilities are when faced with data other than those used to build the models in the first place. Ideally, model validation is done using new data, but acquiring new data is often not possible.

To avoid the worst sort of data-dredging (and its attendant problems for statistical inference) as well as validate my models without having a comparable, completely independent dataset on which to test them, I employed a 'data splitting' strategy (Picard and Berk 1990). I divided the dataset into three subsets, to be used respectively for:

1. developing the models representing a set of alternative hypotheses, a process known as a specification search;
2. testing how well the models work and selecting a 'best' model (or small group of similarly well-performing models); and
3. validating the models, by performing a very simple evaluation of the fragility of the models and any conclusions that might be drawn from them (Aldrich 2006), that is, investigating whether the results using the set of alternative models on another sample (not used to develop the models or undertake model selection) produce similar results.

I took an Information-Theoretic approach to model specification and selection, focusing on the use of Akaike's Information Criterion (AIC), following Burnham and Anderson (2002). Practitioners in various scientific fields now use the information-theoretic approach, but, so far, few archaeologists have done so (see Beheim and Bell 2011; Manning *et al* 2013b; Eve and Crema 2014). Rather than seek to reject (or not) a particular hypothesis, the information-theoretic approach recognises that all models are only approximations of truth and aims to rank models based on estimates of how much (or how little) information the models lose about truth. The goal is to find a model or set of models that strike the best balance between minimising the complexity of the models and maximising the amount of information they reveal.

It is well-known that increasing the number of explanatory variables in a regression model can appear to increase the explanatory power of the model, even when the explanatory variables are completely unrelated to the response variable (Freedman 1983; Burnham and Anderson 2002, 17-18). The principle of parsimony in statistical modelling is that a model should use the smallest possible number of parameters that adequately represent the data (Burnham and Anderson 2002, 29-35). Parsimonious models, that is, those with the smallest possible number of explanatory variables, are generally preferred.

AIC (Akaike 1973; Burnham and Anderson 2002, 60-64) is a measure of model quality that incorporates a penalty as the number of model parameters increases, following the principle of parsimony. AIC provides an estimate of the relative distance between a model and the unknown, true process that generated the observed data. The absolute size of the AIC statistic for any model is, by itself, unimportant; it is the relative difference in AIC values for two or more models that is useful in assessing relative model quality. In practical terms, one develops a set of candidate models, computes the AIC for each one, and the model or models with the lowest AIC values can be considered the best. AIC differences (referred to as Δ_i) among a set of models are calculated by subtracting the lowest AIC value in the set from the AIC value for each model. When comparing models, differences in AIC values > 3 are usually held to indicate that the models with the higher values are considerably poorer, and models having $\Delta_i > 10$ have essentially no support compared to the 'best' model (Burnham and Anderson 2002, 70-72).

It should be noted that ArcGIS routinely quotes the small-sample corrected version of AIC (abbreviated AICc), which includes a bias-correction term not present in the 'simple' version of AIC (Burnham and Anderson 2002, 66). The use of AICc is preferable when sample sizes are small, especially when relative to the number of parameters in a model. Given the large sample sizes in this study (nearly 11,000 observations in each subset), use of AICc is not strictly speaking necessary, even when employing a fairly large number of explanatory variables. As sample sizes increase, however, the difference between AIC and AICc for a given model becomes negligibly small (Burnham and Anderson 2004, 270). For the analyses presented in this section, I exclusively used AICc, rather than AIC.

Given the considerable number of potential explanatory variables and the consequently huge number of potential combinations of variables, it was necessary to evaluate a large number of different OLS models. I did not, however, make an exhaustive search over every possible model, the 'all possible subsets' approach. The number of possible models is 2^K, where K is the total number of explanatory variables. With 20 potential explanatory variables, the number of models to run would be over one million. For this study, I considered over 100 different potential explanatory variables, meaning the number of possible models is more than 38 million million million times larger than the total number of observations in the dataset. A brute force approach to model-building – computing every conceivable combination of variables – would theoretically be possible, but the computing power and time required were so great as to make the approach practically unfeasible.

For each response variable, I began the specification search using only the soils data, trying all 27 Soilscape types as individual variables in a model, then collapsing different groups of types into single variables. I explored more than 50 different groups of soil types, based on similarities between the types. I ensured that, in any model, all Soilscapes types were included as variables, either individually or grouped with other types. I chose a set or sets of soils variables that had Δ_i (AICc difference) values <3 for further analysis.

Once I had a reliable enough set of soils variables, I incorporated elevation and the various surface roughness variables, to see which produced the best fit in conjunction with the soils. The aim was to limit the number of variables used later in the specification search process, rather than use all fifteen of the elevation/surface roughness variables. For each response variable, I generated models for the set or sets of soils variables with each of the fifteen elevation/surface roughness variables. I selected for further analysis the elevation and/or surface roughness variables used in those models that had Δ_i values <3.

Having identified the sets of soils and elevation/surface roughness parameters that produced the best models for each response variable, I then generated models using an exhaustive combination of further variables for precipitation, temperature and insolation. I did not, however, use every possible variable for precipitation, temperature and insolation. Instead, I sought to address and examine rigorously a slight incongruity in Williamson's model of the influence of environmental factors on development of differing forms of settlement organisation. Williamson (2003; 2005; 2010; 2013) uses maps of annual temperature and precipitation to illustrate his discussions of the influence of climatic variables. He considers in some detail, however, the importance of seasonal precipitation in relation to timing for both spring and autumn ploughing, as well as for hay-making in late summer, and the role of summer temperature in cereal growth. By examining annual as well as monthly and multi-month averages for precipitation, temperature and insolation, I investigated whether annual or season-specific values for each of the overall types of variables explain a greater amount of variation in the response variables.

Following some preliminary exploratory work, I excluded from further analysis the variables for monthly days of rain ≥ 10.0mm and the amount of direct incoming solar radiation. The climatic variables used are listed in Table 2, and Table 3 sets out the various combinations of soils and other environmental variables I examined in the specification search and model selection process. From each group of models generated from the different combinations of soils and other environmental variables (listed in Table 3), I selected those models that had Δ_i values <3 within their group for further analysis.

The approach taken here treads arguably the fine line between exploratory data analysis leading to multiple hypotheses on the one hand and data dredging on the other. Given the lack of clear, a priori knowledge regarding the influence of environmental factors on variation in settlement organisation, a degree of data dredging is perhaps inevitable. Splitting the data into subsets for model specification, model selection and model validation, and, above all, explicitly acknowledging the analytical process used and the potential uncertainties arising from it hopefully mean that whatever data dredging I have done does not constitute a grievous sin (Burnham and Anderson 2002, 41).

Table 2: Annual and season-specific climatic variables used in the regression analysis

Variable Group	Variable Name	Description
ann p	bio12	Average annual precipitation
spr p		Spring precipitation
	p_3	Precipitation in March
	p_4	Precipitation in April
	p_5	Precipitation in May
	p_av2m34	Two-month average (March-April) for precipitation
	p_av2m45	Two-month average (April-May) for precipitation
	p_av3m345	Three-month average (March-May) for precipitation
atmn p		Autumn precipitation
	p_9	Precipitation in September
	p_10	Precipitation in October
	p_11	Precipitation in November
	p_av2m910	Two-month average (September-October) for precipitation
	p_av2m1011	Two-month average (October-November) for precipitation
	p_av3m91011	Three-month average (September-November) for precipitation
spr RD		Spring days of rain ≥ 1mm
	RD1mm_3	Days in March of rain ≥ 1mm
	RD1mm_4	Days in April of rain ≥ 1mm
	RD1mm_5	Days in May of rain ≥ 1mm
	RD1mm_av2m34	Two-month average (March-April) for days of rain ≥ 1mm
	RD1mm_av2m45	Two-month average (April-May) for days of rain ≥ 1mm
	RD1mm_av3m345	Three-month average (March-May) for days of rain ≥ 1mm
atmn RD		Autumn days of rain ≥ 1mm
	RD1mm_9	Days in September of rain ≥ 1mm
	RD1mm_10	Days in October of rain ≥ 1mm
	RD1mm_11	Days in November of rain ≥ 1mm
	RD1mm_av2m910	Two-month average (September-October) for days of rain ≥ 1mm
	RD1mm_av2m1011	Two-month average (October-November) for days of rain ≥ 1mm
	RD1mm_av3m91011	Three-month average (September-November) for days of rain ≥ 1mm
ann t	bio1	Average annual temperature
sum t		Summer temperature
	t_6	Temperature in June
	t_7	Temperature in July
	t_8	Temperature in August
	t_av2m67	Two-month average (September-October) for temperature
	t_av2m78	Two-month average (September-October) for temperature
	t_av3m678	Three-month average (June-August) for temperature
ann DSR	DSR_av12m	Annual (twelve-month) average for duration of solar radiation
sum DSR		Summer insolation (Duration of Solar Radiation)
	DSR_6	Duration of solar radiation in June
	DSR_7	Duration of solar radiation in July
	DSR_8	Duration of solar radiation in August
	DSR_9	Duration of solar radiation in September
	DSR_av2m67	Two-month average (June-July) for duration of solar radiation
	DSR_av2m78	Two-month average (July-August) for duration of solar radiation
	DSR_av2m89	Two-month average (August-September) for duration of solar radiation
	DSR_av3m678	Three-month average (June-August) for duration of solar radiation
	DSR_av3m789	Three-month average (July-September) for duration of solar radiation

Table 3: Combinations of soils and other environmental variables used in regression analysis

Soils + one other environmental variable
S + ann p
S + spr p
S + atmn p
S + spr RD
S + atmn RD
S + ann t
S + sum t
S + ann DSR
S + sum DSR
S + elv
S + srf rgh

Soils + two other environmental variables
S + ann p + ann t
S + ann p + sum t
S + ann p + ann DSR
S + ann p + sum DSR
S + ann p + elv/srf rgh
S + spr p + atmn p
S + spr p + ann t
S + spr p + sum t
S + spr p + ann DSR
S + spr p + sum DSR
S + spr p + elv/srf rgh
S + atmn p + ann t
S + atmn p + sum t
S + atmn p + ann DSR
S + atmn p + sum DSR
S + atmn p + elv/srf rgh
S + spr RD + atmn RD
S + spr RD + ann t
S + spr RD + sum t
S + spr RD + ann DSR
S + spr RD + sum DSR
S + spr RD + elv/srf rgh
S + atmn RD + ann t
S + atmn RD + sum t
S + atmn RD + ann DSR
S + atmn RD + sum DSR
S + atmn RD + elv/srf rgh
S + ann t + elv/srf rgh
S + sum t + elv/srf rgh
S + ann DSR + elv/srf rgh
S + sum DSR + elv/srf rgh

Soils + three other environmental variables
S + spr p + atmn p + ann t
S + spr p + atmn p + sum t
S + spr p + atmn p + ann DSR
S + spr p + atmn p + sum DSR
S + spr p + atmn p + elv/srf rgh
S + spr RD + atmn RD + ann t
S + spr RD + atmn RD + sum t
S + spr RD + atmn RD + ann DSR
S + spr RD + atmn RD + sum DSR
S + spr RD + atmn RD + elv/srf rgh
S + ann p + ann t + elv/srf rgh
S + ann p + sum t + elv/srf rgh
S + ann p + ann DSR + elv/srf rgh
S + ann p + sum DSR + elv/srf rgh
S + spr p + ann t + elv/srf rgh
S + spr p + sum t + elv/srf rgh
S + spr p + ann DSR + elv/srf rgh
S + spr p + sum DSR + elv/srf rgh
S + atmn p + ann t + elv/srf rgh
S + atmn p + sum t + elv/srf rgh
S + atmn p + ann DSR + elv/srf rgh
S + atmn p + sum DSR + elv/srf rgh
S + spr RD + ann t + elv/srf rgh
S + spr RD + sum t + elv/srf rgh
S + spr RD + ann DSR + elv/srf rgh
S + spr RD + sum DSR + elv/srf rgh
S + atmn RD + ann t + elv/srf rgh
S + atmn RD + sum t + elv/srf rgh
S + atmn RD + ann DSR + elv/srf rgh
S + atmn RD + sum DSR + elv/srf rgh

Soils + four other environmental variables
S + spr p + atmn p + ann t + elv/srf rgh
S + spr p + atmn p + sum t + elv/srf rgh
S + spr p + atmn p + ann DSR + elv/srf rgh
S + spr p + atmn p + sum DSR + elv/srf rgh
S + spr RD + atmn RD + ann t + elv/srf rgh
S + spr RD + atmn RD + sum t + elv/srf rgh
S + spr RD + atmn RD + ann DSR + elv/srf rgh
S + spr RD + atmn RD + sum DSR + elv/srf rgh

S: Soils elv: Elevation srf rgh: Surface roughness

All OLS regression was performed using ArcGIS 10.0 and 10.1.

Results

Model specification, selection and validation

The specification search process outlined above identified three sets of soils variables that performed best, details of which are presented in Table 4. For reference, Appendix 2 lists the Soilscapes types and the combinations of types used in the specification search.

Table 4: Best-performing sets of soils variables identified through specification search

Soils Variable Set	Soilscapes Types/Type Combinations Used
A	3, 6, 8, 9, 13, 17, 18, Combo 32, Combo 33, Combo 35, Combo 52, Combo 53, Combo 57
B	3, 5, 6, 7, 8, 9, 13, 17, 18, Combo 32, Combo 33, Combo 35, Combo 52, Combo 53
C	5, 6, 8, 9, 13, 17, 18, Combo 32, Combo 33, Combo 35, Combo 52, Combo 53, Combo 54

Table 5 summarises the results of the specification searches for the four settlement variables analysed using subset 1. For each response variable, the table sets out the soils variable set(s) and the elevation/surface roughness variables used in the best-performing models. The last column records how many models had within-group Δ_i values <3, identifying them as worth analysing further in the model selection procedure.

Table 5: Summary model specification search results

Response Variable	Soils Variable Set(s) Used	Elevation/Surface Roughness Variables Used	Number of Models Having Within-Group Δ_i <3
DstNclAll	A, B, C	Elevation; VRM (3x3 window)	440
DstNclBCD	A, B, C	Elevation	536
CSS Na2	B	Elevation; Relief (3x3 window); Standard Deviation of Elevation (3x3 window); Standard Deviation of Slope (3x3, 11x11 and 21x21 windows); TRI	123
CSS Nb2	A, B	As for CSS Na2	174

For each response variable, I then ran the models identified in the specification search on subset 2 of the dataset to select the best-performing model(s) overall. To validate or test the fragility of the results, I repeated the process for each response variable using subset 3 of the dataset, allowing comparison of the results from subset 3 and those from subset 2.

The tables on the following pages set out summary information about the ten best-performing models for each response variable and subset. The columns in each table list the model number (in ascending order based on the Δ_i values), the specific variables included in the model (abbreviated to use the soils variable sets listed in Table 4), the adjusted R^2 value, the Δ_i value, and finally the Akaike weight. Adjusted R^2 is the measure of how much variation in the response variable is accounted for by the explanatory variables, adjusted for the complexity of the model, ie, the number of variables used. The Akaike weight is a measure of the weight of evidence in favour of a particular model being the best for the situation, given the data and the set of models computed using those data (Burnham and Anderson 2002, 75). Akaike weights are calculated using the Δ_i values and sum to 1.0 for all the models computed for a given response variable. The higher the Akaike weight, the greater the weight of evidence for a particular model being the best among all the models specified. The Akaike weights provide an useful means of interpreting the relative weight of evidence for each model in the set. While only the top ten models are listed in the tables, all the models in each set were used when calculating the Akaike weights.

Table 6 sets out the results for the model selection procedure for DstNclAll (performed on data subset 2). The best-performing models all included elevation, the temperature in August, and spring and autumn precipitation treated as separate variables (precipitation averaged over April a May and for November, respectively). The Δ_i and Akaike weight values indicate that the three best-performing models are clearly superior to the other models in the model set. There is, however, little to differentiate between the three best models. The Akaike weight for the model using soils set B is noticeably lower when compared to the weights for the top two models. The Akaike weight for the best-performing model, however, is only slightly better than that for the next-best model.

Table 6: Summary results of the ten best performing models for DstNclAll Subset 2

Model	Explanatory Variables	Adj R^2	Δ_i	Akaike Weight
1	Elevation; p_av2m45; p_11; t_8; Soils Set C	0.350	0.00	0.482
2	Elevation; p_av2m45; p_11; t_8; Soils Set A	0.350	0.70	0.340
3	Elevation; p_av2m45; p_11; t_8; Soils Set B	0.350	2.00	0.178
4	Elevation; p_av3m345; p_11; DSR_9; Soils Set C	0.348	32.12	<0.0001
5	Elevation; p_av3m345; p_11; DSR_9; Soils Set A	0.348	32.75	<0.0001
6	Elevation; p_av3m345; p_11; DSR_9; Soils Set B	0.348	33.90	<0.0001
7	Elevation; p_av2m45; p_11; bio1; Soils Set C	0.348	34.40	<0.0001
8	Elevation; p_av2m45; p_11; bio1; Soils Set A	0.348	34.98	<0.0001
9	VRM3x3; p_av2m45; p_11; t_8; Soils Set C	0.348	35.98	<0.0001
10	Elevation; p_av3m345; p_11; DSR_av12m; Soils Set C	0.348	36.37	<0.0001

Results of the validation procedure (running all the models on data subset 3) are presented in Table 7. Comparison of Table 6 and Table 7 reveals that the three best-performing models were the same, but the ranking of the top three models differed slightly between the model selection and validation data subsets. The models using soils set B ranks third in both cases, but the models using soil sets A and C swap places when comparing the model selection and validation data subsets. The models using soils set B are clearly third-best, but comparing the model selection and validation results indicate that the models using soils sets A and C are probably equally good. Given that the difference between soils sets A and C are limited, this outcome is unsurprising.

Table 7: Summary results of the ten best performing models for DstNclAll Subset 3

Model	Explanatory Variables	Adj R^2	Δ_i	Akaike Weight
1	Elevation; p_av2m45; p_11; t_8; Soils Set A	0.349	0.00	0.487
2	Elevation; p_av2m45; p_11; t_8; Soils Set C	0.349	0.81	0.325
3	Elevation; p_av2m45; p_11; t_8; Soils Set B	0.349	1.91	0.188
4	Elevation; p_av2m45; p_11; bio1; Soils Set A	0.347	37.99	<0.0001
5	Elevation; p_av2m45; p_11; bio1; Soils Set C	0.347	38.91	<0.0001
6	Elevation; p_av2m45; p_11; bio1; Soils Set B	0.347	39.83	<0.0001
7	Elevation; p_av3m345; p_11; DSR_9; Soils Set A	0.345	70.95	<0.0001
8	Elevation; p_av3m345; p_11; DSR_9; Soils Set C	0.345	71.31	<0.0001
9	Elevation; p_av3m345; p_11; DSR_9; Soils Set B	0.345	72.89	<0.0001
10	Elevation; p_av3m345; p_11; DSR_av12m; Soils Set A	0.344	73.39	<0.0001

The model selection results for DstNclBCD are shown in Table 8. The Δ_i and Akaike weight values indicate that the top five models are clearly superior to the rest of the set. The best-performing models all included the variables for elevation, the two-month average for precipitation from March to April, and precipitation in November. Rather than temperature, the best-performing models included variables for duration of solar radiation in July and averaged over July to August. The top two models included soil set A, followed by those using soils set C. The model using soils set B performed least well of the top models. The very close Akaike weight values among the top five models indicates that there is substantial uncertainty as to which of the models should be considered the best. The model ranked fifth overall can probably be discounted, but the performance differences between the top four are so small as to make them practically indistinguishable. As with the results for DstNclAll, the differences in the variables included in the best models are minor.

Validation of the model selection outcomes for DstNclBCD produced a very different set of results, as can be seen in Table 9. The top three models based on the validation data subset were markedly better than the rest. All three used elevation, August temperature, and two precipitation variables: the three-month average for March to May and the single month of November. The Akaike weight values indicate that the best-performing model, using soils set A, was noticeably but not overwhelmingly better than the second-ranked model, using soils set B. The best-performing models here are far more similar to those identified for DstNclAll than to those identified in the model selection process for DstNclBCD.

Table 8: Summary results of the ten best performing models for DstNclBCD Subset 2

Model	Explanatory Variables	Adj R^2	Δ_i	Akaike Weight
1	Elevation; p_av2m34; p_11; DSR_av2m78; Soils Set A	0.333	0.00	0.273
2	Elevation; p_av2m34; p_11; DSR_7; Soils Set A	0.333	0.18	0.250
3	Elevation; p_av2m34; p_11; DSR_av2m78; Soils Set C	0.333	0.72	0.191
4	Elevation; p_av2m34; p_11; DSR_7; Soils Set C	0.333	0.96	0.169
5	Elevation; p_av2m34; p_11; DSR_7; Soils Set B	0.333	1.71	0.116
6	Elevation; p_av3m345; p_11; t_8; Soils Set C	0.332	12.99	<0.0001
7	Elevation; p_av3m345; p_11; t_8; Soils Set A	0.332	13.32	<0.0001
8	Elevation; p_av3m345; p_11; t_8; Soils Set B	0.332	14.79	<0.0001
9	Elevation; p_av2m34; p_11; DSR_av12m; Soils Set A	0.332	22.59	<0.0001
10	Elevation; p_av2m34; p_11; DSR_av12m; Soils Set C	0.332	22.81	<0.0001

Table 9: Summary results of the ten best performing models for DstNclBCD Subset 3

Model	Explanatory Variables	Adj R^2	Δ_i	Akaike Weight
1	Elevation; p_av3m345; p_11; t_8; Soils Set A	0.336	0.00	0.544
2	Elevation; p_av3m345; p_11; t_8; Soils Set B	0.336	1.37	0.275
3	Elevation; p_av3m345; p_11; t_8; Soils Set C	0.336	2.21	0.181
4	Elevation; p_av2m34; p_11; DSR_av2m78; Soils Set A	0.335	19.99	<0.0001
5	Elevation; p_av2m34; p_11; DSR_7; Soils Set A	0.335	20.19	<0.0001
6	Elevation; p_av2m34; p_11; DSR_av2m78; Soils Set C	0.335	20.66	<0.0001
7	Elevation; p_av2m34; p_11; DSR_7; Soils Set C	0.335	20.83	<0.0001
8	Elevation; p_av2m34; p_11; DSR_7; Soils Set B	0.335	21.61	<0.0001
9	Elevation; p_av2m34; p_11; Soils Set A	0.334	25.86	<0.0001
10	Elevation; p_av2m34; p_11; bio1; Soils Set A	0.335	26.04	<0.0001

The difference in outcomes between the model selection and validation procedures shows that the model selection results for DstNclBCD are extremely fragile. It is worth noting that the models ranked sixth to eighth in the model selection procedure are the same as the top three models identified in the validation procedure. The top five models found in the selection procedure were ranked fourth to eighth using the validation data subset. There is, therefore, considerable uncertainty as to which of the models ought to be used as a basis for inference. One possible approach in this instance would be to employ formal multi-model inference (Burnham and Anderson 2002, 149-205), that is, using a weighted average of the results of several models. For the moment, that possibility must remain unexplored. There does not appear to be a reliable basis for preferring, say, the three-month average for spring precipitation to the two-month average, or the August temperature variable to the July or July–August insolation variables. What is clear, however, is that the best-performing models for DstNclBCD all include variables for elevation, spring and autumn precipitation as separate variables rather than the annual average, and summer temperature or insolation rather than annual values.

Turning to the results for CSS Na2 and CSS Nb2, summarised below in Table 10 and Table 11 and Table 12 and Table 13 respectively, the model selection and validation outcomes are unequivocal. Models using the variables for elevation, precipitation in March and September, annual temperature, and soils set B performed best. In each case, the Δ_i and Akaike weight values show that the support for these models being the best in their respective sets is overwhelming.

Before looking closely at the best-performing models, it is worth summarising the results presented thus far and making a few general observations. The model specification, selection and validation procedures for the different response variables identified a number of different models that performed well. Model selection and validation for CSS Na2 and CSS Nb2 produced highly consistent results, indicating that one particular model was clearly the best of all those generated for those response variables. Model validation also reproduced closely (though not identically) the model selection results for DstNclAll. These results suggest that inference based on the best-performing models should be reliable. For DstNclBCD, however, there is considerable uncertainty regarding which model or models ought to be considered 'best'.

No one model was selected and validated as best for all four response variables, but some aspects of the range and nature of the variables used in the best-performing models are already apparent. Overall, the best models were those using the largest number of parameters. Following the principle of parsimony, simpler models are generally preferable to more complex ones. Here, however, the Δ_i and Akaike weight values clearly indicated that simpler models, that is, those with fewer explanatory variables, performed dramatically worse than the best models. The performance of models using the surface roughness variables was, on the whole, poorer than that of those using simple elevation. Using separate variables for spring and autumn precipitation rather than a single annual average resulted in better-performing models for all the response variables. Models including the long-term averages for monthly days of rain ≥ 1.0mm generally performed worse than those using the averages for total monthly precipitation. There was less consistency regarding the use of temperature vs insolation and summer vs annual averages for temperature/insolation.

Table 10: Summary results of the ten best performing models for CSS Na2 Subset 2

Model	Explanatory Variables	Adj R^2	Δ_i	Akaike Weight
1	Elevation; p_3; p_9; bio1; Soils Set B	0.281	0.00	1.000
2	Elevation; p_3; p_9; t_8; Soils Set B	0.266	218.16	<0.0001
3	Elevation; RD1mm_av2m45; RD1mm_11; t_6; Soils Set B	0.265	231.89	<0.0001
4	Elevation; p_3; p_av2m910; t_8; Soils Set B	0.265	233.57	<0.0001
5	Elevation; RD1mm_4; RD1mm_11; t_6; Soils Set B	0.265	242.00	<0.0001
6	Elevation; RD1mm_av2m45; RD1mm_11; bio1; Soils Set B	0.265	243.64	<0.0001
7	Elevation; RD1mm_4; RD1mm_av3m91011; bio1; Soils Set B	0.264	250.99	<0.0001
8	RD1mm_av2m45; RD1mm_av3m91011; bio1; Soils Set B	0.264	252.11	<0.0001
9	RD1mm_4; RD1mm_av3m91011; bio1; Soils Set B	0.264	258.42	<0.0001
10	RD1mm_av2m45; RD1mm_11; t_6; Soils Set B	0.263	261.20	<0.0001

Table 11: Summary results of the ten best performing models for CSS Na2 Subset 3

Model	Explanatory Variables	Adj R^2	Δ_i	Akaike Weight
1	Elevation; p_3; p_9; bio1; Soils Set B	0.285	0.00	1.000
2	Elevation; p_3; p_9; t_8; Soils Set B	0.270	231.52	<0.0001
3	Elevation; p_3; p_av2m910; t_8; Soils Set B	0.269	239.55	<0.0001
4	Elevation; RD1mm_av2m45; RD1mm_11; t_6; Soils Set B	0.269	242.34	<0.0001
5	Elevation; RD1mm_av2m45; RD1mm_11; bio1; Soils Set B	0.269	244.43	<0.0001
6	RD1mm_av2m45; RD1mm_av3m91011; bio1; Soils Set B	0.268	250.06	<0.0001
7	Elevation; RD1mm_4; RD1mm_11; t_6; Soils Set B	0.269	259.98	<0.0001
8	Elevation;RD1mm_4; RD1mm_av3m91011; bio1; Soils Set B	0.267	263.42	<0.0001
9	RD1mm_av2m45; RD1mm_11; t_6; Soils Set B	0.267	264.40	<0.0001
10	RD1mm_4; RD1mm_av3m91011; bio1; Soils Set B	0.267	265.25	<0.0001

Table 12: Summary results of the ten best performing models for CSS Nb2 Subset 2

Model	Explanatory Variables	Adj R^2	Δ_i	Akaike Weight
1	Elevation; p_3; p_9; bio1; Soils Set B	0.290	0.00	1.000
2	Elevation; RD1mm_av2m45; RD1mm_11; t_6; Soils Set B	0.275	227.97	<0.0001
3	Elevation; RD1mm_av2m45; RD1mm_11; t_6; Soils Set A	0.275	231.01	<0.0001
4	Elevation; RD1mm_av2m45; RD1mm_11; bio1; Soils Set B	0.275	232.28	<0.0001
5	Elevation; RD1mm_av2m45; RD1mm_11; bio1; Soils Set A	0.275	236.52	<0.0001
6	SDSlp21x21; RD1mm_av2m45; RD1mm_av3m91011; bio1; Soils Set B	0.275	237.37	<0.0001
7	SDSlp21x21; RD1mm_av2m45; RD1mm_11; bio1; Soils Set B	0.275	239.25	<0.0001
8	RD1mm_av2m45; RD1mm_av3m91011; bio1; Soils Set B	0.275	239.62	<0.0001
9	RD1mm_av2m45; RD1mm_11; bio1; Soils Set B	0.275	240.94	<0.0001
10	SDSlp21x21; RD1mm_av2m45; RD1mm_av3m91011; bio1; Soils Set A	0.275	241.28	<0.0001

Table 13 Summary results of the ten best performing models for CSS Nb2 Subset 3

Model	Explanatory Variables	Adj R^2	Δ_i	Akaike Weight
1	Elevation; p_3; p_9; bio1; Soils Set B	0.295	0.00	1.000
2	Elevation; RD1mm_av2m45; RD1mm_11; bio1; Soils Set B	0.279	241.16	<0.0001
3	SDSlp21x21; RD1mm_av2m45; RD1mm_av3m91011; bio1; Soils Set B	0.279	241.81	<0.0001
4	SDSlp21x21; RD1mm_av2m45; RD1mm_11; bio1; Soils Set B	0.279	242.27	<0.0001
5	Elevation; RD1mm_av2m45; RD1mm_11; bio1; Soils Set A	0.279	244.11	<0.0001
6	SDSlp21x21; RD1mm_av2m45; RD1mm_av3m91011; bio1; Soils Set A	0.279	244.39	<0.0001
7	SDSlp21x21; RD1mm_av2m45; RD1mm_11; bio1; Soils Set A	0.279	244.92	<0.0001
8	Elevation; RD1mm_av2m45; RD1mm_11; t_6; Soils Set B	0.279	245.32	<0.0001
9	RD1mm_av2m45; RD1mm_11; bio1; Soils Set B	0.279	245.83	<0.0001
10	RD1mm_av2m45; RD1mm_av3m91011; bio1; Soils Set B	0.279	246.29	<0.0001

Analysis of best-performing models

Having selected a single best model or small set of best models for each response variable from the set of candidate models, it is possible to examine the best models in detail. The tables in Appendix 3 set out the results for each model, as well as the results of a set of diagnostic tests, used to assess aspects of the model's quality. A brief explanation of the diagnostic tests and what they measure will aid interpretation of the individual models (see Esri 2013).

The Koenker (BP) statistic and associated probability (or p-) value evaluate whether the model standard errors are biased, meaning that the robust standard errors, t-statistic and associated p-value should be consulted for each variable, rather than the conventional standard errors, t-statistic and p-value. The test assesses whether the relationships between the explanatory and response variables are homoscedastic and stationary, that is, that they are numerically and spatially consistent. A relationship is homoscedastic if the variation between each explanatory variable and the values predicted by the regression does not change depending on the magnitude of the explanatory variable values. The opposite, heteroscedasticity, means that a model may predict well for some values of the dependent variable, say, the low values, but becomes unreliable for other values. A relationship is stationary if the variation between each explanatory variable and the values predicted by the regression does not change over geographical space. A non-stationary relationship is one where a model may predict well in one region of the study area but not in another. The Koenker (BP) statistics were significant in all cases, meaning the models' standard errors are biased, so only the robust values are reported for the individual explanatory variables in each model. These results suggest that all the models may have problems of heteroscedasticity, non-stationarity, or both.

The Joint Wald statistic and associated p-value assess the statistical significance of a model as a whole, accounting for the biased standard errors. The null hypothesis for this test is that the explanatory variables do not contribute to the model's ability to predict values of the response variable. All the models were shown to be significant, indicating that, taken as a whole, the explanatory variables in each model do contribute to predicting values for the response variables.

The Jarque-Bera statistic and associated p-value indicate whether a model's residuals deviate from a normal distribution. Model residuals are the difference between the actual value recorded for each observation and the value predicted by the regression. Large positive residuals indicate where the regression analysis has underestimated the value for the response variable, and large negative residuals indicate an overestimate. Statistically significant Jarque-Bera results are often a sign that a model is mis-specified, that is, there are important explanatory variables missing from the model (Burt *et al* 2009, 509). Nonlinear relationships between the explanatory and response variables, influential outliers in the set of observations, or strong heteroscedasticity in the relationships can also produce statistically significant Jarque-Bera results. All the models had statistically significant results for this test, suggesting that the models likely have a range of possible issues relating to mis-specification, nonlinear or heteroscedastic relationships, or influential outliers.

The final diagnostic is the Moran's Index, referred to as Moran's I, test for spatial autocorrelation in the model residuals. Spatial autocorrelation is the degree to which data values correlate with each other depending on their spatial location (Cliff and Ord 1973; Mitchell 2005, 104–5). Moran's I indicates if the data are randomly distributed, or whether there is positive or negative spatial autocorrelation, ie, whether the data are spatially random, clustered or dispersed. It has long been recognised that linear regression models may be mis-specified if the residuals exhibit spatial autocorrelation (Cliff and Ord 1981; Legendre 1993; Getis 2010) because the method assumes independently distributed errors. Spatial autocorrelation in the residuals suggests that: a) the model may not be doing as good a job explaining variation in the response variable as the R^2 measure implies; b) there is information in the residuals about the behaviour of the response variable that the model does not capture; and c) some explanatory variables may not actually be significant at the levels the diagnostic tests suggest (Haining 2003, 352).

I applied the Moran's I test to the residuals for each model, using a range of distance bands (starting at 2,828.5m and adding 2,828.5m for each successive band up to a maximum of 16,971.0m,), weighting the values based on the inverse Euclidean distance between the observations, and using row standardisation. I chose the distance bands based on the distance between the centre of a 2 x 2km grid cell and that of its diagonally adjacent neighbours. Using the initial distance band, some of the observations had no neighbours, potentially invalidating the test results. The tables in Appendix 3 report the Moran's I results based on a distance band of 5,657.0m. A positive z-score for the Moran's I test indicates clustering of values, while a negative z-score indicates that the values are dispersed. The associated p-values indicate the statistical significance of the Moran's I results. All the models showed statistically significant clustering in their residuals, using all of the distance bands. These results also point to problems of mis-specification in all the models.

Figure 5 and Figure 6 show maps of the residuals from four of the best-performing models, one for each response variable. The maps depict the residual values using standard deviations, that is, they show how far from the mean the residual for each observation is. The residual maps for the other models using DstNclAll and DstNclBCD as response variables are very similar to those presented in Figure 5. As noted previously, large positive residuals show where the regression has underestimated the value for the response variable, and large negative residuals indicate an overestimate. As would be expected from the Moran's I statistics, numerous clusters of both positive and negative residuals can be seen in all the maps.

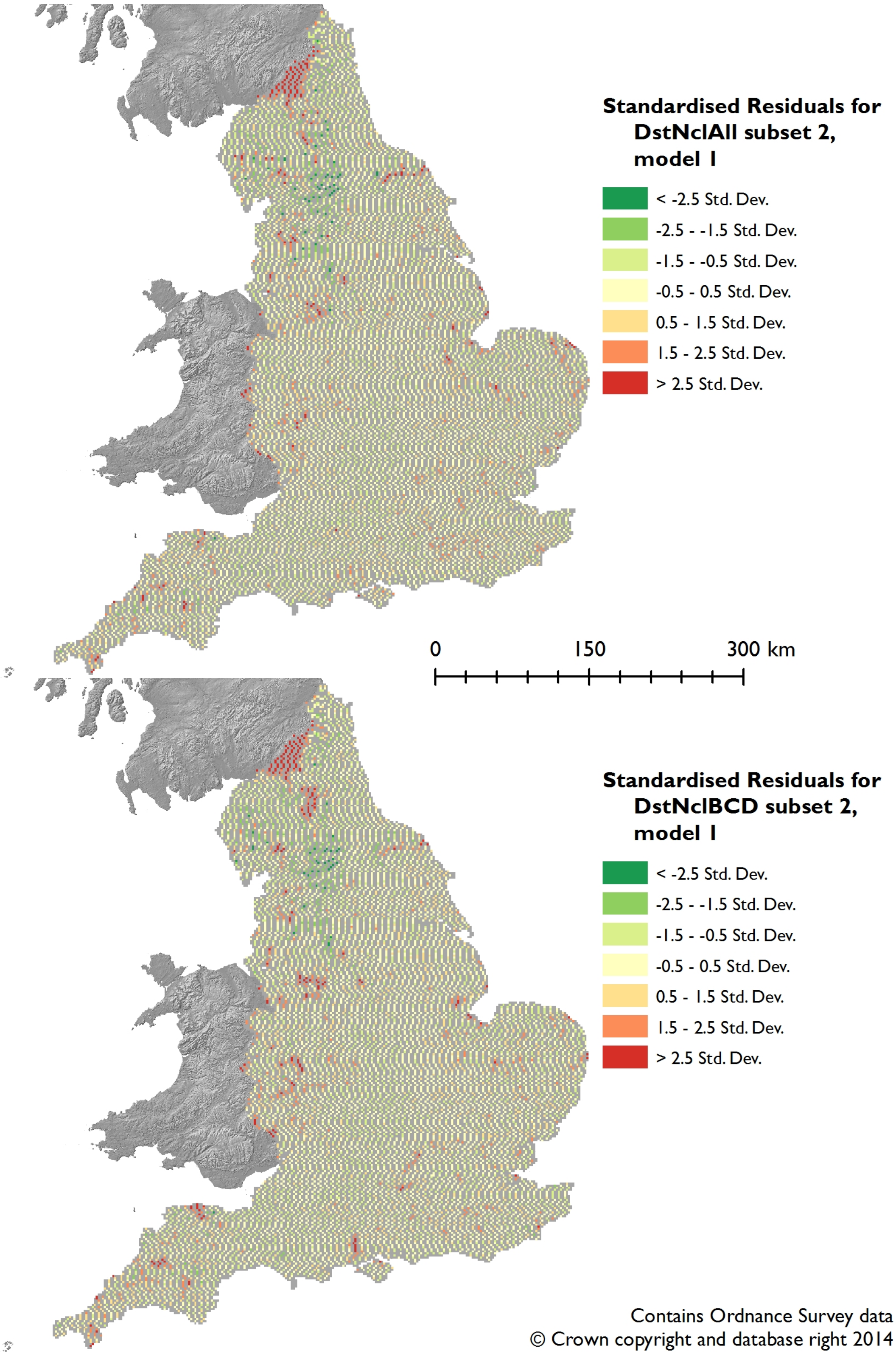


Figure 5: Maps of standardised regression residuals from DstNclAll subset 2, model 1 (top) and DstNclBCD subset 2, model 1 (bottom)

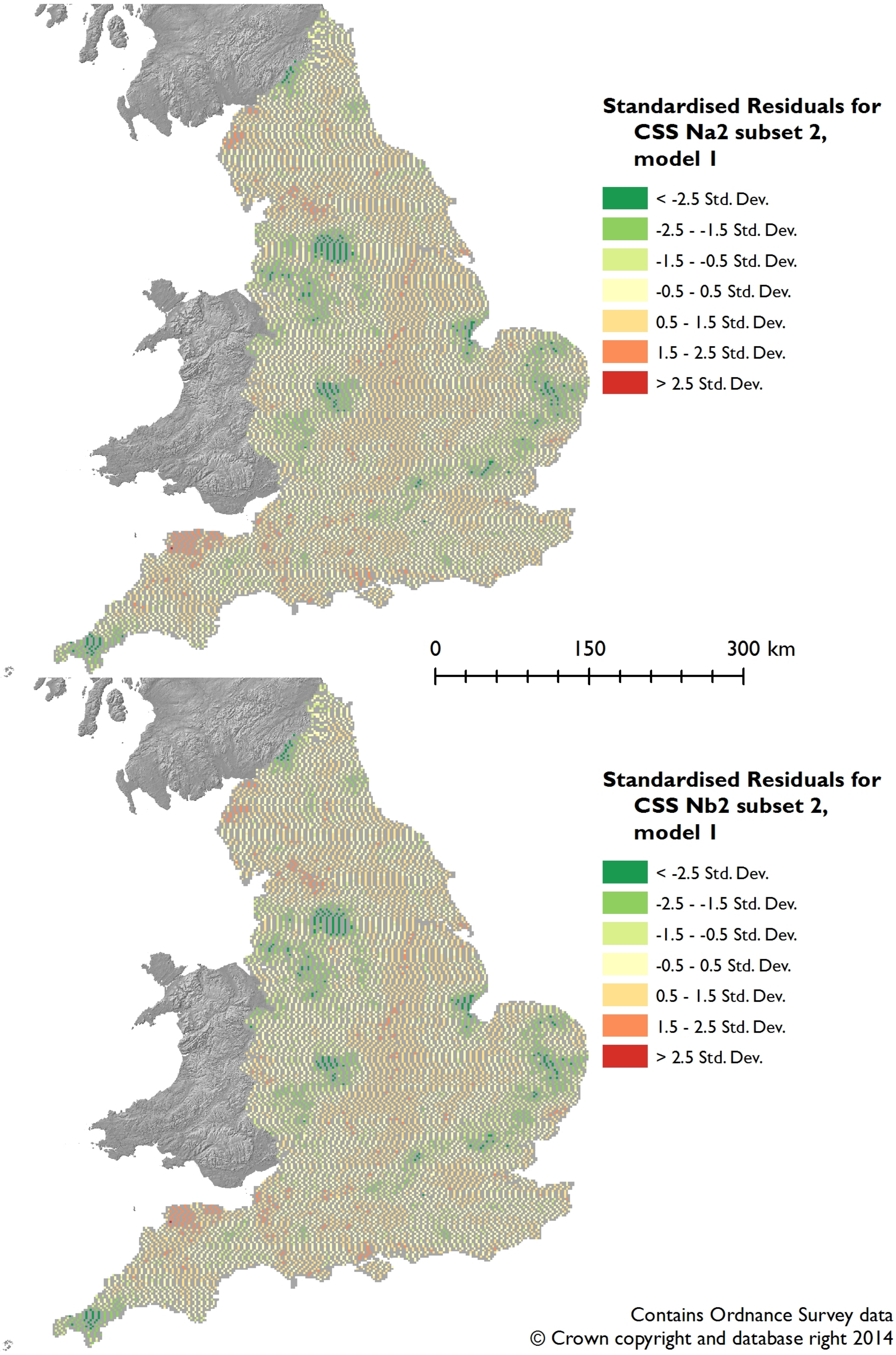


Figure 6: Maps of standardised regression residuals from CSS Na2 subset 2, model 1 (top) and CSS Nb2 subset 2, model 1 (bottom)

Overall, the diagnostic tests indicate that all the models have a number of issues that potentially render them unreliable and make any inferences based on them suspect. It was, of course, clear from the outset that a model (or set of similarly well-supported models) based only on environmental variables would be incomplete – there are a range of 'cultural' factors which almost certainly affected spatial variation in forms of settlement organisation. That being the case, it is no surprise that the best models identified through the model specification, selection and validation procedures are, very likely, still mis-specified. This is not to say that the results of the various models are definitely incorrect. Rather, there is considerable uncertainty as to how reliable the results are. The discussion of aspects of individual models that follows must be read with this very strong caveat in mind.

The tables in Appendix 3 report the coefficient, robust standard errors, t-statistics and associated p-values, and the Variance Inflation Factor (VIF) for each explanatory variable in a model. The coefficient indicates the type and strength of the relationships between the explanatory variable and the model's response variable. The coefficient sign indicates whether the relationship is positive or negative. A positive relationship is when an increase in the value of the explanatory variable results in an increase in the value of the response variable. A negative relationship is when an increase in the explanatory variable results in a decrease in the response variable. The coefficient indicates how much the value of the response variable would be expected to change for every one-unit change in the associated explanatory variable, with all other variables held constant. For example, in the models for DstNclAll, a coefficient for the explanatory variable t_8 (August temperature) of -203.0 means that for every one-degree increase in temperature, it is expected that the distance to all nucleations would decrease by 203m, holding all other variables constant. The t-statistic and associated p-values for each variable show whether the variable coefficient is statistically significant, that is, how likely is it that the coefficient is effectively equal to zero and thus not contributing to explaining variation in the response variable.

VIF measures collinearity (sometimes called multicollinearity) among the explanatory variables. Collinearity is the degree to which the explanatory variables are correlated with each other (Belsley 1991, 19-39). Collinearity can be problematic because OLS regression assumes that the explanatory variables are independent of one another. Collinearity between explanatory variables may indicate that one or the other of the variables in question is effectively redundant. Various rules of thumb are often used to suggest a maximum threshold for acceptable VIF values (Esri 2013; O'Brien 2007, note 2), usually ranging from 4 to 10. The most commonly suggested remedy for models with variables showing 'unacceptably' high VIF values is simply to remove the variable from the model. O'Brien (2007), however, notes the danger of sticking blindly to such rules of thumb for VIF values. He advocates a pragmatic approach of using VIF values for individual coefficients as a means of evaluating whether collinearity might be contributing to other issues, such as a coefficient being insignificant or of the 'wrong' sign (ie, negative when one would expect it to be positive). Dropping one or more variables simply because they produce high VIF values may do more harm than good.

Collinearity between the separate spring and autumn precipitation variables in the selected models is to be expected. Generally, areas that receive a great deal of rain in the spring also receive a great deal of rain in the autumn. As it turns out, the inclusion of separate precipitation variables for spring and autumn produced high VIF values (that is, higher than the common 'rule of thumb' threshold of 10) for those variables in all the models except those for DstNclAll.

Similarly, given the nature of the soils data, a certain amount of collinearity is unavoidable. For example, if the proportion of a grid square covered by Soilscape type 18 is very high, the proportions covered by all the other Soilscape types will, by definition, be very low or zero. If two soil types each occupy nearly 50 per cent of a grid square, then the values for all the other soils types will, again, be very low or zero. Further, the soil types are not randomly distributed across the country – they are recorded as interdigitated but nonetheless discrete blocks. Soilscape type 18 covers 20 per cent of England, so it might be expected that there would be a higher degree of collinearity between it and the other soils variables than would be the case for less spatially extensive soil types. The VIF for the Soilscape type 18 variable is notably high (12.56 to 12.98) in all the selected models.

Having noted the various diagnostics for the models, it is now possible to examine the behaviour and degree of influence of individual explanatory variables in the different models. The elevation variable coefficient is positive in all models for DstNclAll and DstNclBCD, and negative for the CSS models. This behaviour is consistent because the CSS variables score locations close to nucleations higher than those far away from nucleations. As might be expected, the models indicate that as elevation increases, and assuming the remaining variables are held constant, the distance to nucleation increases very slightly. Put another way, the higher the elevation, the lower the degree of 'nucleatedness'. These results agree well with the long-recognised infrequency of nucleated settlement in upland areas (cf Lowerre 2010, table 2).

The summer temperature and insolation variables in the models for DstNclAll and DstNclBCD all show negative coefficients, suggesting that, all other things being equal, warmer and sunnier conditions resulted in lower distances to nucleations. This seems reasonable, given that some of the coolest and shadiest parts of England (eg, the Pennines and Lake District, the North York Moors, and Dartmoor and Exmoor in the south west) are known to be areas where nucleations are less common than elsewhere. That having been said, considerable areas of eastern and south-east England and much of lowland Cornwall are notably warm and sunny but are characterised by less frequent nucleated settlement than much of central England. Taking England as a whole, however, the coefficients suggest a trend for warmer and sunnier regions to have more nucleated settlement. For the CSS models, the coefficient for the annual temperature variable (bio1) is negative, indicating that, *ceteris paribus*, higher annual temperature produced a lower level of nucleated settlement. Here, the models appear to reflect the trend toward more dispersed settlement in the warmer areas of south-west, eastern and south-east England.

The contrasting results for the temperature/insolation coefficients for the models for nucleation variables on the one hand and the combined settlement score variables on the other should not be seen as contradictory. Rather, the differing results reflect the nature of the response variables. Looking only at nucleations, the models suggest that there was a tendency for nucleated settlement to avoid areas with lower summer temperature or insolation and prefer areas with higher summer temperature or insolation, relative to the country as a whole. There are, however, some relatively cool regions, for example the Eden valley in Cumbria and the Northumberland plain, that are characterised by low levels of dispersed settlement and frequent nucleations. Looking at nucleated and dispersed settlement together, the effect of higher annual temperature increasing the level of dispersed settlement (and correspondingly lowering the amount of nucleated settlement) appears to have been more pronounced than the effect of higher temperature increasing the frequency of nucleated settlement.

Regarding the specific variables identified in the model specification and selection processes, Williamson (2003, 174) notes that hay-making was a time-critical activity and was essential for producing winter fodder for livestock: it is vital, as the saying goes, to make hay while the sun shines. For areas with extensive meadows, Williamson argues it made sense for nucleated settlement to develop, as it facilitated mobilising the large workforces necessary for large-scale hay-making. Based on a range Anglo-Saxon and later medieval sources, July was considered the usual month for hay-making (Hill 1998). It is perhaps unsurprising, then, that, for DstNclBCD at least, models using insolation for July or July and August together performed well in the model specification and selection procedure. August was the traditional month for harvesting cereal crops, a task which also required a large workforce and good weather. The same argument as for hay-making may have applied, and would help explain the inclusion of the variable for August temperature in the models for DstNclAll.

In all of the models for DstNclAll and DstNclBCD, the coefficient for the spring precipitation variable is negative and that for the autumn precipitation variable is positive. For the CSS models, the coefficient signs are reversed. As noted for the elevation variable, this behaviour is consistent, given the manner in which the CSS variables were calculated. Essentially, all the models suggest that, *ceteris paribus*, higher spring precipitation resulted in a greater degree of 'nucleatedness' (that is, a lower distance to the nearest nucleation and lower dispersion scores), but higher autumn precipitation resulted in more dispersion (a higher distance to the nearest nucleation and higher dispersion scores). The alternation of coefficient signs is puzzling.

Williamson (2003, 141-59; 2013, 196-201) highlights the importance of spring and autumn precipitation and the influence the amount and timing of rainfall could have on the 'windows of opportunity' available for ploughing, especially the ploughing of certain types of clay-rich soils. In the early medieval and medieval periods, heavy ploughs and, more importantly, the teams of oxen that drew them, were often shared among multiple peasant households, a practice known as co-aration. Williamson argues it made sense for those households to be in close physical proximity, to be able to mobilise rapidly vital agricultural resources. Following this line of argument, the connection between higher precipitation (and correspondingly smaller windows of opportunity to plough certain soils) and a greater degree of nucleation makes sense, if one assumes that the most important episodes of ploughing occurred in spring. Banham (2010, 183-85), however, suggests that in the Middle Ages, wheat was primarily an autumn-sown crop and barley was usually sown in the spring. Assuming this generalisation is correct, and given the ever-increasing importance of wheat as a cereal crop through the Anglo-Saxon to the medieval periods, according to Williamson's logic, one might have expected greater autumn precipitation to lead to more nucleation rather than less.

Interpreting the coefficients for the spring and autumn precipitation variables is fraught with uncertainty. Banham's generalisation about autumn- vs spring-sown crops could be wrong, but then so could Williamson's argument about co-aration and its effects on settlement organisation. It may be that the alternation in coefficient signs has more to do with the growing cycle of a range of crops, not just wheat, than with the timing of ploughing. If the process or processes by which precipitation affected settlement are heteroscedastic and/or non-stationary, the regression equation might produce incorrect results. Finally, it is also possible that collinearity (reflected in the unusually high VIF values) may be influencing the coefficients. Further work is required to make better sense of these results.

Turning now to the soils variables, Soilscape types 3 (shallow lime-rich soils over chalk or limestone), 5 (freely draining lime-rich loamy soils) and 7 (freely draining slightly acid but base-rich soils) and the soils combinations comprised of them (54 and 57) all have large negative coefficients in the DstNclAll and DstNclBCD models and positive coefficients in the CSS models. These Soilscapes types include the Barrow, Newmarket, Swaffham Prior, Upton and Wantage soil associations, which Williamson (2003, 124-25, 139-40) demonstrates were often farmed using 'sheep-corn husbandry' methods and tended toward more nucleated rather than dispersed settlement.

Soilscape type 6 (freely draining slightly acid loamy soils) is significant in all of the DstNclAll and CSS models, but is only significant in the DstNclBCD models run using data subset 2, not the validation models based on subset 3. The coefficients are negative in the distance to nucleation models (in those models where the variable is significant), but, curiously, are also negative in the CSS models. Soilscape type 6 lies predominantly in areas with higher levels of dispersion, so it may be that while these soils may have contributed to increased nucleation in some areas, they contributed to more dispersed settlement in more areas. The difference in coefficient signs may, again, be due to non-stationarity.

Soilscape types 8 (slightly acid loamy and clayey soils with impeded drainage) and 17 (slowly permeable seasonally wet acid loamy and clayey soils) are insignificant for all the models using distance to nucleations as the response variable. The same variables are significant in the CSS models. That Soilscape types 8 and 17 do not appear to have affected distance to nucleations in a meaningful fashion is somewhat surprising. They include a number of soil associations (Batcombe, Essendon, Flint, Hornbeam 1–3, Oak 1 and Oxpasture) noted by Williamson (2003, 33, 64, 101, 147) as either of low fertility or prone to seasonal waterlogging, characteristics that he argues had a considerable effect on population density and settlement organisation. It is equally surprising that the coefficients for the two variables are negative in the CSS models, that is, they indicate that increasing values for either of the variables contributed to decreasing amounts of nucleation, assuming all other variables were held constant. The results for these variables do not appear to be consistent with Williamson's explanatory model. There are a number of possible reasons for these outcomes. Williamson focuses his arguments on particular soil associations, but the Soilscapes types include more associations than just those which Williamson discusses. The soils data used here may be too aggregated to reflect the subtleties Williamson explores. There may be collinearity issues, though the VIF values for the variables in question are not especially high in any of the models. These results may, however, simply reflect the difference between an analysis encompassing the whole of England and explanations based, as Williamson himself admits, predominantly on evidence from 'the East Midlands, East Anglia, and those parts of the Home Counties lying north of the Thames' (Williamson 2013, 5).

Soilscape type 9 (lime-rich loamy and clayey soils with impeded drainage) is significant in the DstNclAll and DstNclBCD models, but, somewhat surprisingly, is insignificant in the CSS models. This type includes the Evesham 1–3 and Hanslope associations, noted by Williamson (2003, 146-47) as prone to the kinds of waterlogging and puddling issues that may, ultimately, have contributed to the development of nucleated settlement in many areas. The coefficient is negative in the distance to nucleation models, suggesting that greater amounts of Soilscape type 9 soils led to greater nucleation. This result aligns well with Williamson's arguments, but the apparent lack of meaningful influence on the CSS response variable does not. These soils appear to have had a statistically significant effect on nucleation treated in isolation but not on the more complex characterisation of settlement organisation that blends measures of nucleation and dispersion together. Precisely why this might be the case remains unclear.

Higher amounts of the freely draining acid loamy soils over rock making up Soilscapes type 13 appear to contribute to a greater degree of settlement dispersion and lower nucleation, all other things being equal. The coefficients are negative for the CSS models and positive for the distance to nucleation models, and they are significant in all the models.

Soilscape type 18 is significant in all the models and has negative coefficients for the DstNclAll and DstNclBCD response variables, but positive coefficients for the CSS response variables. Keeping all other parameters in the models constant, an increase in the value for Soilscape type 18 produces a lower distance to nucleation value, but it also produces a lower combined settlement score. This soils type includes several of the heavy clay associations (Beccles, Denchworth, Foaggathorpe, Ragdale and Wickham) on which

Williamson notes the timing of cultivation would have been critical (2003, 146-47; Williamson 2010, 139, fig 7.2), leading, ultimately, via the practice of co-aration, to the development of nucleated settlement. The results for the distance to nucleation variables agree neatly with Williamson's explanatory model, but those for the CSS variables are the opposite of what one would expect. As for Soilscapes types 8 and 17, there are a variety of possible explanations for the discrepancy between these results and Williamson's explanatory model, including differences in the level of aggregation in the soils data, problems of collinearity among the soils variables, and differences in the geographical extent of the analyses.

The coastal soils of combination 32 – including saltmarsh (Soilscape type 1), sand dunes (Soilscape type 4) and the loamy and clayey soils of coastal flats with naturally high groundwater (Soilscape type 22) – appear to lead to lower nucleation and increased settlement dispersion. The coefficients for this variable are significant in all the models. Combination 33 (loamy floodplain soils and loamy soils with naturally high groundwater) shows a strong tendency toward a high degree of nucleation. The coefficients are negative (and large) for the distance to nucleation models and positive for the CSS models. This results agrees well with Williamson's arguments about meadow and its importance to the development of nucleated settlement (2003, 169-77). Comprised of Soilscapes types 12, 20 and 22, this combination includes the Fladbury, Frome and Thames soil associations, highlighted by Williamson as soils ideal for meadow-making (ibid, 170-71).

The variable for soils combination 35 (sandy and very acid loamy soils, except for sand dunes) is insignificant in the models having DstNclAll or DstNclBCD as the response variable, but is significant for the CSS models. The coefficients in the CSS models are negative, indicating that, *ceteris paribus*, higher values for this soils combination contributes to lower nucleation and higher dispersion. Williamson (2003, 124, 131) notes that the soil associations such as Methwold, Newport and Worlington included in soils combination 35 were only sparsely settled in the late Saxon and medieval periods. The remaining soil combinations, 52 (upland peaty soils) and 53 (lowland peaty soils plus restored soils), have large, positive coefficients for DstNclAll and DstNclBCD. The coefficients for these combinations are negative in the CSS models, but combination 53 was not significant. It seems clear that peaty soil acted as a repulser to the development of nucleated settlement.

The final aspect of the models to report is how well they actually explain the varying levels of 'nucleatedness', that is, their adjusted R^2 values. As previously noted, adjusted R^2 measures how much variation in the response variable is accounted for by the explanatory variables, adjusted for the complexity of the model. The DstNclAll models all produced adjusted R^2 values of 0.350. The DstNclBCD models generated from data subset 2 had adjusted R^2 values of 0.333, while those from data subset 3 (the validation set) had adjusted R^2 values of 0.336. The adjusted R^2 value for the CSS Na2 model was 0.281 and that for the CSS Nb2 model was 0.290. The adjusted R^2 values for the best-performing models are not especially high, suggesting they explain at most about one-third of the variation in the distance to nucleation response variables, and slightly less than that for the CSS response variables. Given the issues of spatial autocorrelation in the model residuals, there is a distinct possibility that the adjusted R^2 values are actually somewhat inflated.

Discussion

What, then, does the extended OLS regression analysis reveal about the relationships between environmental factors and historic settlement organisation? The model specification, selection and validation procedures found a number of different models that performed well. No single set of explanatory variables worked well for all the response variables, but more complex models performed unequivocally better than simpler models. There is effectively no uncertainty arising from model selection and validation for CSS Na2 and CSS Nb2: one particular model was clearly the best of all those generated for those response variables. Model selection and validation produced slightly less certain results for DstNclAll. Two of the three best-performing models appear equally effective, and the last of the top three models is consistently in third place. There is little doubt that the models in question are the best-performing in the model sets, though numerous issues raised by the individual model diagnostics must also be taken into account.

There is, however, considerable uncertainty regarding which model or models might be considered the best for DstNclBCD. This uncertainty is noteworthy given that it was the category B, C and D nucleations on which Roberts and Wrathmell particularly focused when developing their characterisation of England's rural settlement into provinces, sub-provinces and local regions (Roberts and Wrathmell 2000, 11, 15-16). Even without considering the issues raised by the individual models' diagnostics, it is far from clear which of the fitted models provides the best basis for inference. The high level of model selection uncertainty in the analysis means that confidence in any interpretations of the effects of environmental factors on this key measure of rural settlement organisation must be low. This is not to say that the environmental factors investigated here did not influence the variation in distance to category B, C and D nucleations. Rather, it is difficult to be certain in this instance precisely which constellation of variables best approximates the actual processes by which environmental factors affected settlement organisation.

Diagnostics indicate that all the models suffer from various issues, making inferences based on them potentially suspect. Given that 'cultural' factors were excluded from the analysis, it is no great surprise that even the best models are probably still mis-specified. It cannot be said that the results of the various models are incorrect, but their reliability is open to question.

Some coefficients for individual variables in particular models appear to agree well with Williamson's explanatory model, but not all. The differing coefficient signs for spring and autumn precipitation are especially perplexing, and further work is necessary to explain these results. Some of the apparent discrepancies between Williamson's explanatory model and the results presented here regarding the soils variables may relate to the level of aggregation in the soils data. Williamson focuses his arguments on individual soil associations, while the Soilscape types and combinations of types used here are amalgamations of multiple soil associations. It is possible that regression models constructed using data compiled at the level of soil associations might perform better than those discussed here. It may also be the case that Williamson's explanatory model breaks down when applied to the whole of England, rather than just the area on which he has concentrated.

In the models using distance to nucleation as the response variable, there is a strong tendency toward underestimating the high values, that is, the models perform poorly when trying to predict values at locations very far from the nearest nucleations. Many, though not all, such locations are in upland areas, as well as a substantial area along the Anglo-Scottish border. The models also do not predict well the distance to nucleations in valleys in upland areas, where the models appear consistently to overestimate low values. In the CSS models, there is a very strong tendency to overestimate the low values. Large negative residuals clearly cluster in areas where the CSS values show low levels of nucleation and high levels of dispersion. There are some areas of large positive residuals, for example the north-west Cumbrian coast, the Yorkshire Dales, and Exmoor and the area around Barnstaple in north Devon. The models' propensity to underestimate high values is not nearly so marked as that to overestimate low values.

The best-performing models explain, at most, about one-quarter to one-third of the variation in the settlement response variables. Given the issues of spatial autocorrelation in the model residuals, it is likely that the adjusted R^2 values are somewhat inflated. Authors such as Williamson and Lambourne do not, of course, state in quantitative terms how much influence they believe environmental factors might have had on rural settlement organisation. In the absence of similar, quantitatively-based studies, it is difficult to ascertain whether the results set out here should be considered good, bad or indifferent. What does seem clear, however, is that there is considerable variation in the settlement variables that is not explained by the environmental variables. Environmental factors do appear to have had demonstrable, quantifiable effects on variation in the organisation of rural settlement, but overall, other factors, not investigated here, appear to have had more effect.

A range of potential remedies exists for the kinds of model issues highlighted by the various diagnostic tests, and further work could explore these possibilities. Such remedies include transformations of the response and explanatory variables, the use of spatial trend and interaction variables, and possibly using Geographically Weighted Regression (GWR) (Fotheringham *et al* 2002), which could address potential issues of non-stationarity. The analyses presented here could also be undertaken using coarser-resolution grid cells. The spatial processes by which the environmental factors influenced nucleatedness may not have operated, or at least may not be reliably detected, at the resolution examined here. It is usually preferable to analyse data at the highest possible resolution, but aggregating the data into larger grid cells might, in this case, pay dividends. At the same time, soils data compiled at the level of associations, rather than the more aggregated Soilscape types, might also serve to explain better the relationships between soils and settlement. It is also possible that some of the relationships between environmental factors and settlement organisation are not linear. Non-linear regression, or other analytical techniques that do not make as rigid assumptions as does OLS regression, might reveal more information about the relationships between environmental factors and the measures of settlement organisation used in this study.

As noted at the beginning of this section, OLS regression is a non-spatial method: it does not take incorporate the location in geographic space of observations for the response and explanatory variables into the analysis. Some, though, not all, of the issues identified by the model diagnostics likely relate to the application of a non-spatial method to spatial data. There are, however, regression methods that do incorporate spatial effects. I explore and apply some of these methods to the environment and settlement data in the next section.

Spatial Regression

Spatial regression is a general term for a variety of methods for specifying, estimating and performing diagnostic checks on regression models that explicitly include the effects of spatial autocorrelation in the variables in the calculations (Anselin 2009). Spatial regression is a form of autoregressive modelling, that is, where the value of a variable for one observation is related to the value for the same variable at neighbouring observations (Anselin 1988, 33).

Method

Before describing spatial regression methods in detail, it is worth discussing briefly the general nature of the kinds of factors or processes that produce the kinds of spatial patterning under investigation here. This discussion is useful for understanding how different types of spatial regression models incorporate space into their specifications. The factors and processes leading to spatial patterning are of two basic types: exogenous and endogenous (Fortin and Dale 2005, 6-10). Exogenous factors or processes are those that act from outside or independently of the phenomenon or variable of interest. Endogenous factors or processes are those internal or inherent to the phenomenon or variable of interest. Two examples from ecology help illustrate the different processes. The dispersal of seeds from a plant is an endogenous process that affects the spatial distribution of that type of plant, while the combined effects of topography, soil character, temperature and rainfall demonstrate exogenous processes that affect the spatial distribution of the same plant.

The factors under consideration here are primarily exogenous – the goal is to understand how environmental factors outwith settlement organisation itself affected the distribution of settlement organisation. Of course, there almost certainly will have been some endogenous effects in some regions which, over time, may have led to changes in settlement organisation. For example, people living in dispersed settlements in one area might have noticed the relatively greater success or prosperity of people in a neighbouring area who had (for whatever reasons) begun to live in nucleated settlements. Using diagnostic checks on different forms of spatial regression models, it is possible to explore whether endogenous or exogenous factors appear to have had the most pronounced effects on the different settlement response variables.

Spatial regression incorporates the effects of spatial autocorrelation into regression model specifications in two main ways:

- spatial lag dependence, where spatial autocorrelation in the response variable is modelled
- spatial error dependence, where spatial autocorrelation in the regression error term is modelled

Spatial lag dependence assumes that the autoregressive process occurs *only* in the response variable. A spatial lag model would be most appropriate for explaining a pattern where it is expected that the process driving the observed spatial dependence in the response variable is endogenous. A spatial lag model can also help to address problems of spatial autocorrelation when the spatial scale of the phenomenon being studied and the scale at which it is measured do not match. Spatial error dependence is most likely encountered when spatial autocorrelation is not fully explained by the explanatory variables included in a regression model, for example when important, spatially-structured variables are missing from the model specification. A spatial error model would be most appropriate for explaining patterns where the explanatory variables do not fully capture the exogenous process(es) being modelled (Anselin and Bera 1998, 247-49; Anselin 2002, 248-49, 253; Haining 2003, 313-14; Kissling and Carl 2007, 61).

Spatial dependence is incorporated into the analyses by defining a set of neighbours for each observation, using what is known as a spatial weights matrix. A spatial weights matrix is an N x N matrix (where N is the number of observations in the data set) representing the relationship between each observation and every other observation (Dubin 2009). If two observations are considered neighbours, the value in the matrix will be a positive number; otherwise the matrix value is zero. The accepted convention is that an observation is not its own neighbour, so the main diagonal of the matrix will consist entirely of zeroes. The weighting scheme for those observations that are neighbours can be either discrete or continuous. In a discrete scheme, the value in the matrix for a pair neighbours will be one, indicating simply that two observations are neighbours. In a continuous scheme, the value will indicate the strength of the relationship according to some numeric scale.

There are various ways in which the relationships between observations can be defined, that is, the manner by which one decides whether two observations should or should not be considered neighbours. For polygon spatial data, contiguity or adjacency between polygons can be used. Other discrete weighting schemes appropriate for both point and polygon data include using a chosen number of nearest neighbours and using all observations within a specified distance limit. Commonly used continuous schemes include using weights that are inversely related to the distance separating observations (ie, the greater the distance, the lower the weight) and weights that are a negative exponent of the distance between observations (eg, the weight equals the square of the inverse distance between two observations).

A key problem in estimating and evaluating spatial regression models is choosing an appropriate spatial weights matrix (Anselin 1988, 176-77). Often, it is far from clear exactly how the phenomena under investigation interact over space. Unless the nature of the spatial interactions in the data is very clear, best practice involves estimating models using multiple spatial weights matrices and evaluating which model type and weights matrix produces the best-fitting results (Anselin 2002, 259).

The spatial lag model uses what is known as a spatially lagged response variable. For each observation, a weighted average value of the response variable for that observation's neighbours (as defined by the spatial weights matrix) is calculated and included in the model as an additional explanatory variable. The spatial error model applies the spatial lag (again, calculated for each observation's neighbours as defined by the spatial weights matrix) to the estimated regression errors. In effect, the spatial lag model aims to control for the spatial autocorrelation in the response variable itself, allowing proper interpretation of the effect of the explanatory variables. The spatial error model aims to control for the spatially autocorrelated errors in the model, treating them as a nuisance, and again enabling a more accurate interpretation of the effect of the explanatory variables on the response variable than is possible using a non-spatial OLS model (Anselin and Bera 1998, 246-51).

I investigated only the best-performing models identified in the model specification, selection and validation procedures using non-spatial OLS regression, that is, those described in the tables in Appendix 3. Due to the complexities of interpreting the diagnostics and the longer computing times required relative to non-spatial OLS, it was not possible to undertake a comparably extensive specification search using spatial regression methods. This approach – performing model specification and selection using non-spatial OLS regression, then investigating only the selected models using other regression methods – may be imperfect, but is widely used in practice (Dormann 2007, 135). The question, of course, is which type of model to use, spatial lag or spatial error? Given the focus on exogenous factors here, the simplest approach would be to fit only spatial error models. To do so would, however, ignore the possibility that endogenous processes might have had a substantive and detectable effect on the measures of settlement organisation under investigation.

There are a set of diagnostics tests set out by Anselin *et al* (1996) that can be used to assess which type of spatial model (if any) would be preferable to non-spatial OLS. The tests, known as Lagrange Multiplier or LM tests, examine whether there is statistically significant spatial lag or spatial error dependence in the OLS regression residuals. The LM-Lag test compares against a null hypothesis that there is no spatial dependence in the lagged response variable parameter. The LM-Error test compares against a null hypothesis that there is no spatial dependence in the lagged error parameter. There is, of course, the possibility that there may be both spatial lag and spatial error dependence in a given model, which will bias the standard test results and make them unreliable. The robust tests take into account the local presence of dependence of the opposite form. That is, the robust LM-Lag test examines whether there is significant lag dependence given the local presence of error dependence, and the robust LM-Error test examines whether there is significant error dependence given the local presence of lag dependence.

Anselin (2005, 198-200) helpfully summarises the step-by-step interpretation of these tests as a decision rule. First, the non-spatial OLS model is run and the various diagnostics statistics computed. If the standard (that is, non-robust) LM-Lag and LM-Error tests are both insignificant, then the OLS results can be considered trustworthy, and there is probably no need to compute a spatial model. If one LM test is significant and the other is not, then the spatial model corresponding to the significant test should be estimated. That is, if LM-Error is significant and LM-Lag is not, then one would estimate a spatial error model, and vice versa. If both the standard tests are significant, then comparison of the robust test results is necessary. If one is significant and the other not, or if one result is orders of magnitude more significant than the other (eg, $p < 0.00001$ compared to $p < 0.01$), then the model corresponding to the (most) significant test should be estimated. When both tests are highly significant, the model corresponding to the largest test statistic can be estimated. In this last situation, however, it is probable that there are still considerable mis-specification problems, and caution is necessary. Neither form of spatial model can 'cure' all the potential problems in regression analysis of real-world data when important explanatory variables are missing from the models.

I performed all the spatial regression analyses using the freely available software GeoDA (Anselin *et al* 2006), the most recent version of which is 1.6.6 (downloadable from https://geodacenter.asu.edu/software/downloads). Both the spatial lag and spatial error regression models are estimated using the maximum likelihood approach, rather than OLS (Anselin 1988, 57-65). It should be noted that GeoDA reports the standard version of AIC, that is, not the small-sample-corrected version. As previously noted, for large sample sizes, the difference between AIC and AICc for a given model is negligible. Where applicable, for the analyses presented in this section, I used AIC rather than AICc.

Results

Spatial Regression Model Type Selection

I used three spatial weights matrices, employing a discrete weighting scheme based on three different, progressively larger, cut-off distances: 5,657.1m (Weights Matrix 1), 11,314.1m (Weights Matrix 2) and 16,971.1m (Weights Matrix 3). In effect, these spatial weights matrices define increasingly large areas for the neighbourhood around each observation. For each observation, all other observations closer than the cut-off distance are treated as neighbours, and those further away are not. The cut-off distances are based on multiples of the distance between the centre of a 2 x 2km grid cell and that of its diagonally adjacent neighbours, ensuring that all observations had at least one neighbour (cf the distance bands used in the Moran's I test of OLS residuals above).

For each of the models described in Appendix 3, I re-ran the OLS models using GeoDA, calculating the LM test diagnostics using each of the three spatial weights matrices. Both the standard and the robust versions of the LM-Lag and LM-Error test statistics were significant in every instance. Following the decision rule outlined above, the question then becomes, for each model, which of the robust test statistics is larger, robust LM-Lag or robust LM-Error? Table 14 sets out the robust LM-Lag and robust LM-Error statistics for each model and weights matrix.

For DstNclAll, for all three models, the model selection decision rule indicates that spatial lag models would be preferred when using weights matrix 1, but spatial error models would be preferred if using weights matrix 2 or 3. For most of the models having DstNclBCD as the response variable, the model selection decision rule suggests that spatial error models would be preferred when using all three weights matrices. The Robust LM-Lag and Robust LM-Error statistic values were, however, very similar for the models based on data subset 2; the differences were more pronounced for the models based on data subset 3. For CSS Na2 and CSS Nb2, the model selection decision rule indicates that spatial lag models would be preferred when using weights matrix 1, but spatial error models would be preferred if using weights matrix 2 or 3.

Given these results, for all the sets of parameters set out in Appendix 3, I estimated both spatial lag and spatial error models using weights matrix 1, but only spatial error models using weights matrices 2 and 3. Bearing in mind the note of caution included in the decision rule set out above, it is unclear how reliable the preference for a spatial lag or spatial error model in these circumstances actually is. Estimating both types of spatial model using weights matrix 1 enables a direct comparison, while recognising that neither type of spatial model will alleviate every potential model specification issue.

Table 14: OLS model diagnostics for spatial regression model selection

Response Variable	Data Subset	Model	Weights Matrix	Robust LM-Lag	Robust LM-Error
DstNclAll	2	1	1	509.8	397.4
DstNclAll	2	1	2	441.7	9298.1
DstNclAll	2	1	3	321.8	20984.9
DstNclAll	2	2	1	510.8	395.3
DstNclAll	2	2	2	443.2	9267.4
DstNclAll	2	2	3	323.8	20926.8
DstNclAll	2	3	1	510.6	395.3
DstNclAll	2	3	2	443.4	9264.3
DstNclAll	2	3	3	324.0	20911.8
DstNclBCD	2	1	1	450.5	452.4
DstNclBCD	2	1	2	394.9	9408.6
DstNclBCD	2	1	3	278.5	20737.3
DstNclBCD	2	2	1	455.8	452.0
DstNclBCD	2	2	2	403.0	9384.5
DstNclBCD	2	2	3	287.8	20619.7
DstNclBCD	2	3	1	450.6	455.9
DstNclBCD	2	3	2	395.4	9447.8
DstNclBCD	2	3	3	278.9	20804.1
DstNclBCD	3	1	1	432.8	475.3
DstNclBCD	3	1	2	380.3	10003.0
DstNclBCD	3	1	3	251.8	22084.2
DstNclBCD	3	2	1	389.3	469.3
DstNclBCD	3	2	2	322.9	9606.4
DstNclBCD	3	2	3	192.9	21224.2
DstNclBCD	3	3	1	384.1	476.0
DstNclBCD	3	3	2	314.2	9704.8
DstNclBCD	3	3	3	185.2	21408.3
CSS Na2	2	1	1	771.2	504.3
CSS Na2	2	1	2	1198.9	20275.0
CSS Na2	2	1	3	1270.3	61176.7
CSS Nb2	2	1	1	782.8	520.8
CSS Nb2	2	1	2	1194.3	20409.6
CSS Nb2	2	1	3	1253.4	60793.6

Analysis of estimated models

Table 15 and Table 16 summarise the results and diagnostics for, respectively, the spatial lag and spatial error models. The first three columns in the tables record the response variable, data subset and model number (as listed in the tables in Appendix 3), and the fourth column indicates which spatial weights matrix was used. The next column, labelled R^2, provides a measure of the goodness-of-fit for each model. This is actually a so-called pseudo-R^2 and is not, strictly speaking, directly comparable with the R^2 values reported for the non-spatial OLS models (Anselin 1988, 243-45; Anselin 2005, 207, 218).

The sixth and seventh columns set out the extent to which the spatial regression models improve on the non-spatial OLS models using two other measures of model performance, the log likelihood and AIC. In general, for the log likelihood, the higher the value, the better the fit of the model; for AIC (as previously noted), the lower the value, the better the fit. The sixth column records how much higher the log likelihood value for the spatial model is compared to the log likelihood value for the comparable non-spatial OLS model. The seventh column records the value obtained by subtracting the AIC value for the spatial model from the AIC value for the non-spatial OLS model. Here, the higher the value, the greater the improvement in AIC.

The last three columns present statistics from three different diagnostic tests for each spatial model: the Wald, likelihood ratio (LR) and LM-Lag or LM-Error tests (Anselin 1988, 65-72). Taken together, the results of these tests are useful for assessing whether the models are well-specified. Very simply, it is expected in a well-specified model that the values of the Wald, LR and LM (Lag or Error, as appropriate) tests should compare as follows: Wald ≥ LR ≥ LM (Lag or Error) (ibid, 72-73). When the test results do not conform to that pattern, it is likely that the model in question still has specification issues, even after incorporating spatial effects.

Table 15: Summary results and diagnostics for spatial lag models

Response Variable	Data Subset	Model	Weights Matrix	R^2	Log Likelihood improve	AIC improve	Wald	LR	LM-Lag
DstNclAll	2	1	1	0.720	3802.6	7603.0	14089.7	7605.2	10253.4
DstNclAll	2	2	1	0.720	3804.2	7606.0	14113.4	7608.4	10256.3
DstNclAll	2	3	1	0.720	3804.0	7606.0	14113.4	7608.0	10256.0
DstNclBCD	2	1	1	0.743	4328.2	8655.0	18784.0	8656.4	11393.6
DstNclBCD	2	2	1	0.743	4329.1	8656.0	18782.1	8658.2	11386.4
DstNclBCD	2	3	1	0.743	4327.9	8654.0	18767.6	8655.8	11391.6
DstNclBCD	3	1	1	0.748	4407.0	8812.0	19282.7	8793.9	11623.5
DstNclBCD	3	2	1	0.748	4397.5	8791.0	19274.2	8794.2	11665.2
DstNclBCD	3	3	1	0.748	4393.4	8785.0	19259.1	8786.7	11655.9
CSS Na2	2	1	1	0.928	11216.2	22430.4	125406.5	22432.4	23869.1
CSS Nb2	2	1	1	0.928	11093.8	22185.7	125543.3	22187.6	23592.8

Table 16: Summary results and diagnostics for spatial error models

Response Variable	Data Subset	Model	Weights Matrix	R^2	Log Likelihood improve	AIC improve	Wald	LR	LM-Error
DstNclAll	2	1	1	0.721	3730.9	7462.0	16822.1	7461.7	10141.0
DstNclAll	2	1	2	0.651	3066.3	6133.0	24964.0	6132.6	25469.5
DstNclAll	2	1	3	0.613	2636.0	5272.0	115396.1	5272.0	35878.6
DstNclAll	2	2	1	0.722	3732.3	7464.0	16874.0	7464.7	10141.1
DstNclAll	2	2	2	0.651	3069.4	6138.0	25185.7	6138.8	25479.9
DstNclAll	2	2	3	0.613	2638.6	5277.0	116964.0	5277.3	35904.4
DstNclAll	2	3	1	0.722	3732.5	7465.0	16867.8	7465.0	10140.7
DstNclAll	2	3	2	0.651	3069.1	6138.0	25199.4	6138.2	25468.2
DstNclAll	2	3	3	0.980	523.5	1049.0	761721.0	32063.6	35876.1
DstNclBCD	2	1	1	0.747	4290.7	8582.0	22781.0	8581.3	11395.5
DstNclBCD	2	1	2	0.646	3132.6	6266.0	26227.3	6265.2	26146.5
DstNclBCD	2	1	3	0.595	2538.1	5077.0	87336.9	5076.1	35018.8
DstNclBCD	2	2	1	0.747	4290.5	8581.0	22857.6	8581.0	11382.7
DstNclBCD	2	2	2	0.645	3109.9	6222.0	29991.1	6433.2	26105.2
DstNclBCD	2	2	3	0.593	2512.8	5028.0	107523.7	5239.1	34948.7
DstNclBCD	2	3	1	0.747	4290.8	8581.0	22753.1	8581.7	11396.9
DstNclBCD	2	3	2	0.646	3131.8	6263.0	26127.5	6263.7	26151.6
DstNclBCD	2	3	3	0.595	2537.1	5074.0	86526.5	5074.2	35034.5
DstNclBCD	3	1	1	0.751	4359.9	8720.0	22922.2	8699.8	11666.1
DstNclBCD	3	1	2	0.657	3270.9	6542.0	32231.3	6541.8	27486.6
DstNclBCD	3	1	3	0.604	2647.6	5295.0	116917.2	5295.0	36845.6
DstNclBCD	3	2	1	0.751	4350.2	8699.0	22921.4	8699.8	11745.2
DstNclBCD	3	2	2	0.657	3262.6	6525.0	36114.4	6525.0	27784.1
DstNclBCD	3	2	3	0.604	2637.5	5275.0	134835.8	5275.0	37343.2
DstNclBCD	3	3	1	0.750	4346.4	8693.0	22821.1	8692.7	11747.8
DstNclBCD	3	3	2	0.657	3258.7	6517.0	35480.5	6517.4	27790.9
DstNclBCD	3	3	3	0.604	2633.2	5266.0	130736.6	5266.3	37342.7
CSS Na2	2	1	1	0.929	11174.1	22348.1	153888.3	22348.1	23602.1
CSS Na2	2	1	2	0.880	9360.1	18720.0	1769253.8	18720.1	83013.5
CSS Na2	2	1	3	0.816	7271.2	14542.3	355639250.6	14542.4	136956.2
CSS Nb2	2	1	1	0.928	11045.8	22091.6	155817.1	22091.6	23330.8
CSS Nb2	2	1	2	0.876	9109.6	18219.2	1440578.5	18219.1	81075.6
CSS Nb2	2	1	3	0.812	7084.1	14168.2	69643815.1	14168.1	133173.0

The first overarching point that is immediately apparent from examination of the tables is that all the spatial models show dramatically improved performance compared to the corresponding non-spatial OLS models. Recalling that differences in AIC values > 10 indicate that the model with the higher AIC value has essentially no support compared to the model with the lower value (Burnham and Anderson 2002, 70-72), it is clear that the spatial models are overwhelmingly better than the corresponding non-spatial versions.

Comparing the AIC values for the models estimated using spatial weights matrix 1, it also seems clear that the spatial lag models performed dramatically better than the spatial error models. Among the error models, the larger the neighbourhood defined by the weights matrices, the poorer the performance of the model, again based on comparison of the AIC values. The final general observation to be made based on the results shown in Table 16 and Table 15 is that it is probable that all the spatial models still have mis-specification problems. None of the models have Wald, LR and LM test statistics that match the expected pattern (Wald ≥ LR ≥ LM).

Given the poorer performance of the spatial error models using weights matrices defining larger neighbourhoods, I will concentrate in the remainder of this section only on the models estimated using weights matrix 1. The tables in Appendix 4 present the results for the spatial lag and spatial error models estimated using weights matrix 1, as well as the results of the Breusch-Pagan test for heteroscedasticity in the models' error terms. This test is used to assess each model's quality. The results of the Breusch-Pagan test were significant for all the spatial models. Issues of heteroscedasticity – variation between the explanatory variables and the values predicted by the regression change depending on the magnitude of the explanatory variable values – remain in all the models.

Similar to the results for the non-spatial OLS regression models, the tables report the coefficient, standard errors, z-values and associated p-values for each explanatory variable in a model. Again, the coefficient indicates the type and strength of the relationship between the explanatory variable and the model's response variable. The coefficient estimates – assuming all other variables are held constant – how much the value of the response variable would be expected to change for every one-unit change in the associated explanatory variable. The z-value and associated p-values for each variable show whether the variable coefficient is statistically significant, that is, how likely is it that the coefficient is effectively zero and therefore not contributing to explaining variation in the response variable.

The two types of spatial model each include the autoregressive parameters as an 'extra' coefficient in addition to the explanatory variables as used in the non-spatial OLS models. The spatial lag models include the spatially lagged response variable, reported in the tables as 'W_' plus the name of the response variable. In the spatial error models, lambda is the label given to the coefficient for the spatially lagged estimated regression errors. The autoregressive parameters are significant in all the models, regardless of type. The autoregressive coefficients are quite large in all the models, indicating – as would be expected – that there is considerable spatial autocorrelation in, respectively, the lagged response variables and in the models' errors. The extremely high values for the autoregressive coefficients indicate that much of the 'explaining' going on in the models is being done by the autoregressive parameters. In many, though not all, instances, the coefficients for the other explanatory variables are noticeably smaller in the spatial models than in the corresponding non-spatial OLS models.

Elevation is significant in all the models except the spatial error models for CSS Na2 and CSS Nb2. As with the non-spatial OLS models, the coefficients are positive for the distance to nucleation models and negative for the CSS models, though the coefficients are slightly lower than in the non-spatial OLS models. These results support the interpretation that the higher the elevation, the lower the degree of 'nucleatedness'. Why the elevation variable should be insignificant in the spatial error models for CSS Na2 and CSS Nb2 but not in the spatial lag models is unclear. It suggests that, having controlled for missing explanatory variables, elevation did not have a significant effect on settlement organisation when nucleation and dispersion are measured on a combined scale.

The separate spring and autumn precipitation variables are all significant in all the models. The same puzzling switch in coefficient signs as seen in the non-spatial OLS models is also present: negative for spring precipitation but positive for autumn precipitation (or vice versa for the CSS models). For all the DstNclAll spatial models, the coefficients for the spring precipitation variable are smaller than those in the non-spatial OLS models. The coefficients for the autumn precipitation variables are also smaller in the spatial lag models, but noticeably larger in the spatial error models. The same pattern applies to the spatial models estimated for DstNclBCD using data subset 3. In the spatial models for DstNclBCD estimated from data subset 2, the coefficients for both the spring and autumn precipitation variables are smaller than those in the corresponding OLS models. In the CSS models, the spring and autumn precipitation coefficients are smaller than in the OLS models for both types of spatial model.

Generally, the effects of precipitation on the measures of rural settlement organisation appears less pronounced having controlled for spatial autocorrelation in either the response variables or the models' errors. In the models for DstNclAll and those for DstNclBCD estimated from data subset 3 (the specifications of which are nearly identical), the effect of autumn precipitation appears more pronounced having controlled for spatial autocorrelation in the models' error terms. Interpretation of the precipitation variables, as noted in the discussion of the OLS models, remains perplexing. Even having controlled for spatial autocorrelation effects, the regression models may still not be adequate to explain heteroscedastic and/or non-stationary processes by which precipitation affected settlement. As in the OLS models, it is also possible that collinearity may be affecting the coefficients. Again, more work is required to make better sense of these results.

The summer temperature variable (average temperature in August) is significant in the error models for DstNclAll, but not in the lag models. The absolute values of the coefficients in the error models are somewhat larger than those in the corresponding OLS models, suggesting that, having controlled for spatial autocorrelation in the models' error terms, summer temperature had greater effect on nucleation than the OLS models indicated. The summer insolation variables (the duration of solar radiation in July and averaged for July and August) are significant for all the spatial DstNclBCD subset 2 models. The absolute values of the coefficients in the spatial lag models are smaller than those in the OLS models, while those in the spatial error models are slightly larger. The differences in the magnitudes of the coefficients are, however, unlikely to be meaningful. August temperature is not significant for the subset 3 spatial models. Annual temperature is significant for all the CSS spatial models, but the absolute values of the coefficients in the spatial models are smaller than those in the OLS models. Overall, these results suggest that temperature or insolation probably did have substantive effects on rural settlement organisation, but those effects are difficult to summarise for the whole of England. Given the extent to which temperature and insolation vary spatially across the country, there is a distinct possibility that their effects on rural settlement organisation were non-stationary, that is, they varied from place to place.

The most noteworthy results the tables in Appendix 4 reveal are that far fewer of the soils variables are identified as significant in the spatial models as compared to the non-spatial OLS models. Accounting for spatial autocorrelation in either the response variables or the models' error terms suggests that only a few soil types or combinations of types had a substantive effect on the measures of rural settlement organisation.

Soils combination 54 (amalgamating Soilscapes types 3 and 7, shallow lime-rich soils over chalk or limestone and freely draining slightly acid but base-rich soils) is not significant in any of the models in which it appears. Soilscapes type 3 is only significant in the spatial lag models for the CSS variables. This is in sharp contrast to the non-spatial OLS models, where these variables are uniformly highly significant. Soilscapes types 6 (freely draining slightly acid loamy soils), 8 (slightly acid loamy and clayey soils with impeded drainage) and 9 (lime-rich loamy and clayey soils with impeded drainage) are also insignificant in every spatial model. Soilscapes types 6 and 9 are significant in most of the non-spatial models, so the results from the spatial models are also markedly different.

It should be noted that Soilscapes type 8 has a p-value of 0.054 in the spatial error model for CSS Nb2, and has p-values between 0.07 and 0.09 in the other CSS models, so it could be considered significant if the critical value for α is slightly relaxed from the usual cut-off of 0.05. As in the non-spatial CSS models, the coefficients for this soils variable are negative in the spatial models, that is, they indicate that an increase in Soilscapes type 8 contributed to a decrease in the degree of nucleation, assuming all other variables are held constant. The slowly permeable seasonally wet acid loamy and clayey soils of Soilscapes type 17 are also only significant in the two spatial regression models for CSS Nb2, and if the critical value for α is relaxed to 0.10, also in the spatial lag model for CSS Na2. Again, the coefficients for this variables is negative in the CSS models, that is, they indicate that increasing values for Soilscapes type 17 contributed to decreasing amounts of nucleation, assuming all other variables are held constant. These results are similar to those from the non-spatial OLS models in that they do not appear to be consistent with Williamson's explanatory model.

Soilscapes type 13 (freely draining acid loamy soils over rock) is only significant at α = 0.05 in the spatial lag models for DstNclBCD based on data subset 3. This variable is also significant at $\alpha = 0.10$ for the spatial error models for DstNclBCD based on data subset 3. In the models where it is significant, the coefficients for this variable are positive, indicating that, all other things being equal, higher amounts of this type of soil contributed to a lower degree of nucleation of Roberts and Wrathmell's categories B, C and D. That this variable is significant for only a minority of the spatial models also contrasts sharply with the non-spatial OLS results.

Soilscapes type 18 (slowly permeable seasonally wet slightly acid but base-rich loamy and clayey soils) is insignificant in the vast majority of the spatial models, a profoundly different result to that of the non-spatial OLS models. The only spatial model in which this variable is significant at $\alpha = 0.05$ is the spatial error model for CSS Nb2. It is also significant at $\alpha = 0.10$ in the spatial error model for CSS Na2. In both these models, the coefficient sign is negative, indicating that, *ceteris paribus*, as the amount of this soils type increases, the degree of nucleation decreases. Recalling that Soilscapes type 18 includes several of the soil associations central to Williamson's arguments about the importance of co-aration to the development of nucleated settlement, these results are entirely at odds with Williamson's explanatory model. Controlling for spatial autocorrelation in either the response variables or the models' errors appears to suggest that these soils did not contribute in any meaningful fashion to variation in distance to nucleations. Looking at the combined settlement scores, and having accounted for spatial autocorrelation in the model errors, Soilscapes type 18 soils seem to have the opposite effect to what Williamson contends, as was the case with the non-spatial OLS models.

Soils combination 32 (coastal soils) is significant in nearly all the spatial models for all the different response variables. The only exception is in the spatial lag model for CSS Na2. For the distance to nucleation response variables, the coefficients are smaller in the spatial lag models, but larger in the spatial error models when compared to the results of the non-spatial models. For the CSS response variables, the absolute values of the coefficients are smaller in both types of spatial model compared to the OLS models. As was the case with the OLS models, the coefficients are positive in the distance to nucleation models and negative in the CSS models, indicating these soils appear to repel nucleated settlement.

The coefficients for the loamy floodplain soils and loamy soils with naturally high groundwater of soils combination 33, which includes soils noted by Williamson as ideal for creating meadow, are significant and negative for all the spatial models for DstNclAll, as well as for the spatial lag models for DstNclBCD. They are significant and positive in the spatial lag models for the CSS variables. These results generally correspond to those from the non-spatial OLS models, and agree with Williamson's arguments about the importance of meadow to the development of nucleated settlement. It is not clear why this variable is not significant in the spatial error models for DstNclBCD, CSS Na2 and CSS Nb2. It appears that, once the effects of missing variables are taken into account, this variable did not contribute in a substantive way to variation in three of the four measures of rural settlement organisation under investigation.

Soils type 5 (freely draining lime-rich loamy soils) is significant in most of the models in which it appears, as is soils combination 57, which is comprised of Soilscapes types 5 and 7 (freely draining slightly acid but base-rich soils). These are the soils Williamson shows were often farmed according to 'sheep-corn husbandry' methods and where settlement gravitated more toward nucleated than dispersed forms. The exceptions are the spatial lag and error models 2 and 3 for DstNclBCD subset 3 and the spatial error models for the CSS variables. As was the case with the non-spatial OLS models, the coefficients in the spatial models are negative for the distance to nucleation variables and positive for the CSS variables. The coefficients in the spatial models are smaller in magnitude compared to those in the corresponding OLS models, but overall these results agree with Williamson's explanatory model. It is noteworthy that Soilscapes type 3 – another of the types which fits with Williamson's 'sheep-corn husbandry' model – is only significant in the spatial lag CSS models. It appears that, in most cases, controlling for spatial autocorrelation in the either the response variable or the models' error terms indicates that Soilscapes type 3 did not substantively contribute to variation in settlement organisation.

Soils combinations 52 (upland peaty soils) and 53 (lowland peaty soils plus restored soils) are significant in all the distance to nucleation models, but not in the CSS models. Like the OLS models, the coefficients for these variables in the DstNclAll and DstNclBCD models are positive and large, though not as large as in the OLS models. As might be expected, peaty soils appear to have repulsed the creation of nucleated settlement. These variables do not appear to have had a meaningful impact on settlement organisation when measures of nucleation and dispersion are blended together. One possible explanation is that pockets of dispersed settlement developed in small areas of firmer soils interdigitated through areas of peaty soils. It must also be remembered that Roberts and Wrathmell's depiction of rural settlement is based on maps drawn after the Fenland regions around the Wash had been drained, a long-term process that had profound effects on settlement and land-use in the region (Darby 1956).

Discussion

How, then, does the application of spatial regression techniques refine understanding of the relationships between environmental factors and historic settlement organisation? There is clear evidence of both spatial lag and spatial error dependence in the OLS residuals. This suggests that both endogenous processes and spatially structured missing variables contributed substantively to variation in the measures of rural settlement organisation, in addition to the environmental variables included in the various models. That variables not included in the non-spatial OLS models influenced variation in settlement organisation is entirely unsurprising. The apparent importance of endogenous effects, eg, the diffusion of nucleation as an approach to settlement organisation, was less expected.

After estimating both spatial lag and spatial error models using a range of spatial weights, it is clear that the spatial models perform far better than the corresponding non-spatial versions. Both the log likelihood and AIC measures of goodness-of-fit show dramatic improvements in fit for all the models, indicating that incorporating spatial effects into the modelling analysis is valuable. Comparing the spatial error models using three different weights matrices, those estimated using weights matrix 1, defining the smallest neighbourhood, performed best, based on the log likelihood and AIC measures.

Comparing the models estimated using weights matrix 1, the spatial lag models all show a noticeable improvement of fit over both the non-spatial and spatial error models. It is tempting to infer that endogenous effects had a greater effect on variation in settlement organisation than did any missing variables. Given the continued mis-specification problems in all the models, however, this point must remain open to question. It is possible that a form of diffusion process influenced the development of rural settlement organisation over time, a simple example of which would be a process where those in dispersed settlements chose to mimic more successful neighbours who had (for whatever reasons) chosen to live and work in nucleated settlements. Without comparable settlement data for multiple time periods, however, there is no way to test rigorously whether, how and where such a diffusion process might have worked.

Even though the diagnostic criteria suggest that the spatial lag models offer a better fit to the data, it still seems more likely that the spatial error models provide the best explanation of the variation in measures of rural settlement organisation (cf Sparks and Sparks 2009, 478-79). This is not to deny the possibility that endogenous processes may have affected variation in rural settlement organisation. Given that it was clear from the outset that a range of variables would not be included in the models (the 'cultural' factors), and that such variables are likely to have been spatially structured, the method best suited to address issues of missing explanatory variables seems most appropriate.

Looking at the various environmental variables in turn, elevation and the climatic variables generally continue to be significant variables in the spatial models, and the curious switch in coefficient signs between the spring and autumn precipitation variables is present as well. Accounting for spatial autocorrelation in either the response variables or the models' error terms renders insignificant many of the variables for soil types or combinations of types. Many of the variables corresponding to soils types that Williamson argues strongly influenced the development of nucleated settlement are not significant in the spatial models. Upland and lowland peaty soils, coastal soils, floodplain soils amenable to the creation of meadow, and some soils associated with 'sheep-corn' husbandry appear to have had significant effects on distance to nucleation, results which generally agree with aspects of Williamson's explanatory model. But most clayey soils with impeded drainage and some of the lighter soils Williamson argues encouraged nucleation do not appear to have had meaningful effects on the measures of nucleation used here. Looking at the combined settlement score response variables, some of the soils types produce coefficients opposite to what might be expected according to Williamson's explanatory model. As noted in the discussion of the OLS results, the disagreements between these results and Williamson's explanatory model may be due to differences in the soils data used, or that a model developed to explain evidence from one region fails to account for variation across the country as a whole.

The very large coefficients for the autoregressive parameters in all the spatial models confirm the presence of high levels of spatial autocorrelation in the lagged response variables and in the models' errors. These coefficient values suggest that the autoregressive parameters account for much of the variation in the response variables. Generally, once one accounts for the spatial autocorrelation produced either by some form of diffusion process or by missing explanatory variables, the effects of the other explanatory variables are noticeably reduced compared to the corresponding non-spatial OLS models. These results echo, and arguably state even more forcefully, the conclusion drawn from the non-spatial OLS results: much of the variation in the measures of settlement organisation is not explained by the environmental variables. Some environmental factors did have demonstrable, quantifiable effects on variation in the organisation of rural settlement, but other factors seem to have had far greater influence.

There is, of course, considerable scope for further work on this topic using spatial regression methods. It is worth noting that the spatial regression models used here assume that the intensity and direction of the spatial autocorrelation (in either the lagged response variable or in the model errors) are uniform, that is, they are isotropic (Fortin and Dale 2005, 10-11). Future work could explore the extent to which spatial autocorrelation in the data is isotropic or not. Additional variables accounting for trends in the data, transformations of the response and explanatory variables, interaction variables and other weights matrices could all be investigated. In addition to those applied here, there is a range of other methods available that take spatial autocorrelation into account when performing regression analysis, including methods intended to control for both spatial lag and spatial error dependence (Kelejian and Prucha 2007; Dormann *et al* 2007; Bini *et al* 2009). Further work could explore the application such methods to the environment and settlement data.

Unsupervised Classification/Clustering Analysis

Using the sets of variables identified in the OLS specification search and model selection procedures and examined further using spatial regression, I used unsupervised classification (Lillesand *et al* 2008, 568-572; Conolly and Lake 2006, 147-8) to find locations with similar values for each of the different variables and group them together (cf. Lowerre 2011, 34-5). This technique is, in a sense, an automated version of Roberts and Wrathmell's own division of England into settlement provinces, sub-provinces and local regions. The results provide alternative, national-scale characterisations of historic settlement organisation as it relates to the physical environment, which can be compared to Roberts and Wrathmell's own delineation.

Method

There are numerous approaches to data clustering (Jain 2010), and a range of issues specific to clustering or regionalisation of spatial data (Haining 2003, 199-206; Duque *et al* 2007). Because of its wide applicability, ease of use and ready availability in Esri's ArcGIS software, I used unsupervised classification, employing the ISODATA algorithm (Lillesand *et al* 2008, 568-572; Ball and Hall 1965) as implemented in Esri's Spatial Analyst extension (Esri 2011c). ISODATA unsupervised classification is a non-spatial approach to clustering (Haining 2003, 200-201). The clusters are based only on the similarity of the multivariate attribute values; the locations of the observations in geographic space are not considered. When the clustered units are mapped, the number of individual regions is usually much larger than the number of clusters.

One of the main issues with many forms of data clustering – ISODATA unsupervised classification among them – is that the analyst must choose the number of clusters into which the data will be grouped without reliable prior knowledge of how many clusters are actually present in the dataset. There are numerous methods for determining the 'ideal' number of clusters in a dataset (eg, Milligan and Cooper 1985; Krzanowski and Lai 1988; Tibshirani *et al* 2001; Sugar and James 2003). The basic aim is to find a number of clusters in the data such that the amount of variation within each cluster is minimised and the amount of variation between the clusters is maximised. Due to its reliability and ease of computation, I used the 'variance ratio criterion' developed by Calinski and Harabasz (1974), often known as the CH Index.

As when calculating the Combined Settlement Scores, I standardised the values for all the input variables to a scale of 0.0 to 1.0 by subtracting the minimum from each value and then dividing by the value range (Milligan and Cooper 1988, 185). I performed the ISODATA unsupervised classification on the combinations of variables using a range of numbers of clusters or classes (K), from a minimum of 3 to a maximum of 50. The maximum number of classes the ISODATA algorithm could 'find' varied for each set of input layers. I then calculated the CH Index for each value of K. The value K with the absolute or local maximum value for the CH Index, or for which there is a rapid increase in the value compared to adjacent K, is taken to be the best number of classes. Where there are multiple local maxima in the CH Index values, the lowest of the corresponding values of K is taken to be the best number of classes (Calinski and Harabasz 1974, 12).

I used two permutations of the data to generate the classified settlement layers. For the first, I used the single variables DstNclAll, DstNclBCD, CSS Na2 and CSS Nb2. For the second, I used separate layers for the different nucleation classes, dispersion scores and hamlet counts as inputs to the classification procedure. That is, to obtain a classified layer based on distances to all nucleations, I used five separate layers – one for the standardised distance to each nucleation category. To produce a classified layer comparable to CSS Nb2, I used individual layers recording the standardised distances to category B, C and D nucleations together with layers for standardised dispersion scores and hamlet counts.

Results

CH Index values for the unsupervised classification of distance to all nucleations (DstNclAll) are shown in the top of Figure 7. An arrow indicates the first local maximum in the CH Index values, at K = 9, which can be considered the best number of classes for this variable. The bottom of Figure 7 is a map of the clusters produced by the unsupervised classification using K = 9. Across much of England, the mapped clusters are highly fragmented and do not coalesce into discrete regions. Unsupervised classification of distance to B, C and D category nucleations (DstNclBCD) produced a very similar outcome. This is almost certainly a result of ISODATA unsupervised classification being a non-spatial clustering method. The results for classifying the single distance to nucleation variables DstNclAll and DstNclBCD were clearly unenlightening, so I undertook no further analysis using these variables.

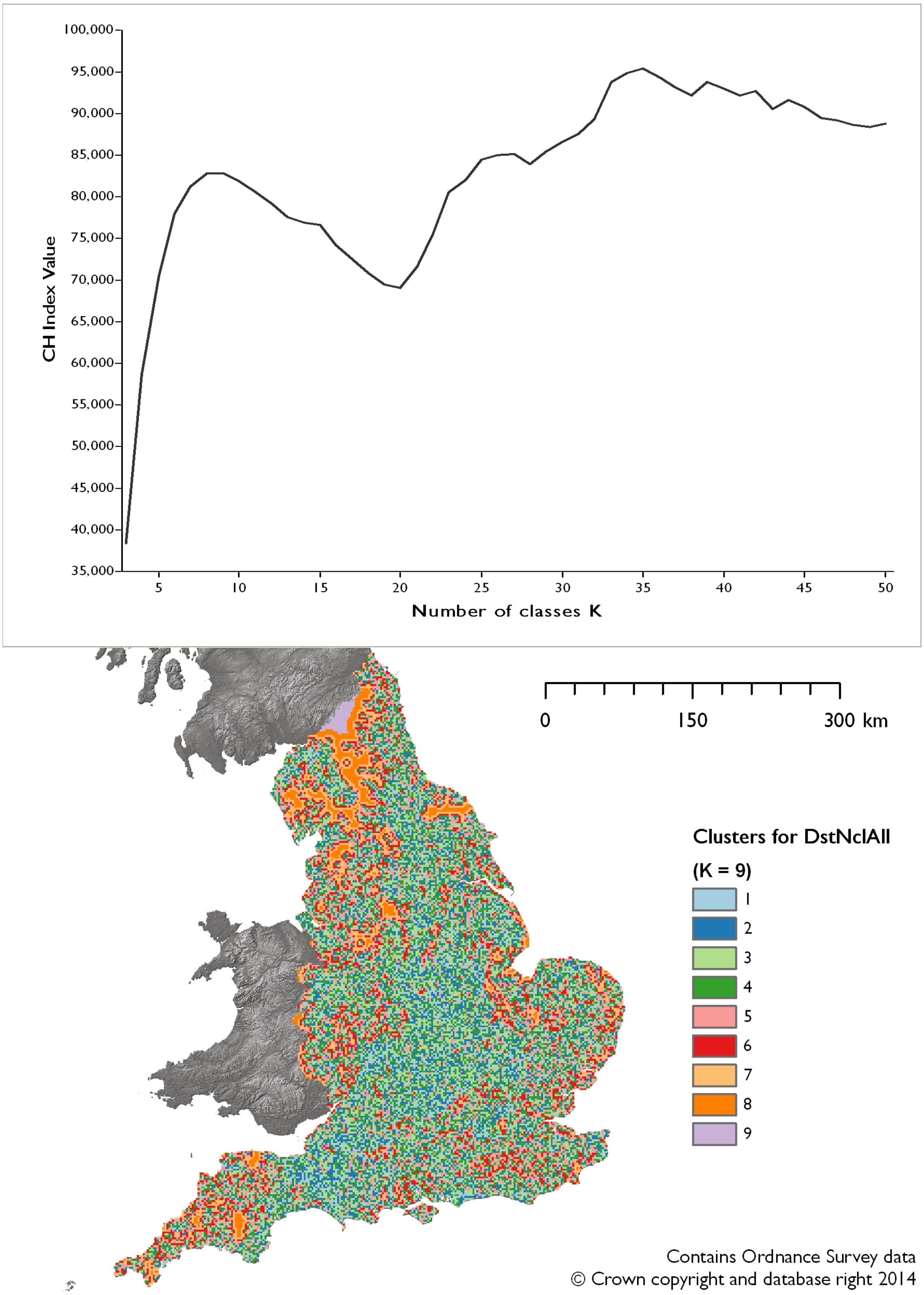


Figure 7: CH Index values for unsupervised classification of distance to all nucleations (top) and map of resulting clusters for K = 9 (bottom)

Figure 8 shows the CH Index values for clustering distance to nucleation categories A–E and B–D, as well as for distance to nucleation categories A–E and B–D when combined with dispersion scores and hamlet counts, taking each nucleation category and measure of settlement dispersion as an individual input layer in the classification process. Here, the ideal number of clusters for each set of variables was open to some interpretation. Arrows point out the CH Index values indicating the values of K which I took to indicate the best numbers of clusters. For distance to nucleation categories A–E and B–D and for distance to nucleation categories A–E combined with dispersion scores and hamlet counts, more than one value of K appeared reasonable.

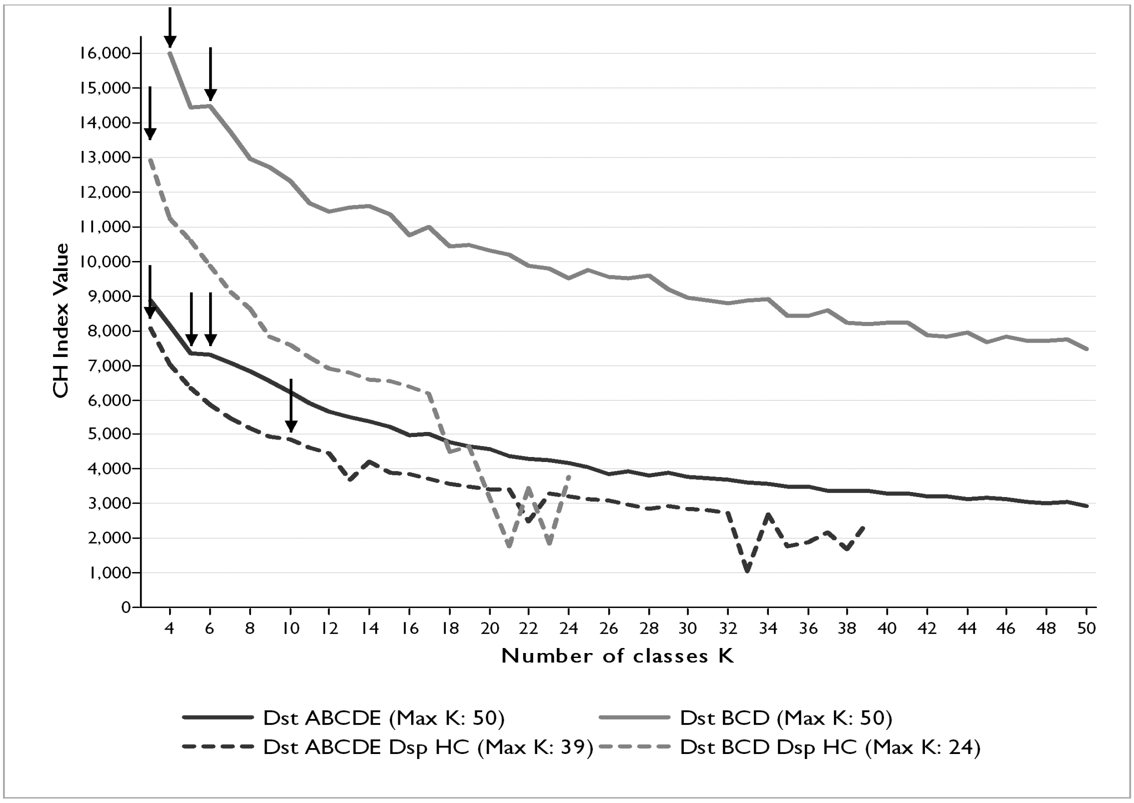


Figure 8: CH Index values for unsupervised classification of distance to nucleation categories A–E and B–D, with and without dispersion cores and hamlet counts

The CH Index results for CSS Na2 and Nb2 were even more difficult to interpret. As can be seen in Figure 9, the CH Index values peak only at very high values of K (25 or 26). The resulting maps of clustered polygons for CSS Na2 and Nb2 where K = 25 or 26 are extremely fragmented, as many of the clusters do not resolve spatially into coherent blocks. I inspected output maps of clustered polygons for CSS Na2 and Nb2 for a range of values of K, comparing the maps of clusters with plots of the raw CSS values, finally deciding on K = 5 for both CSS Na2 and Nb2.

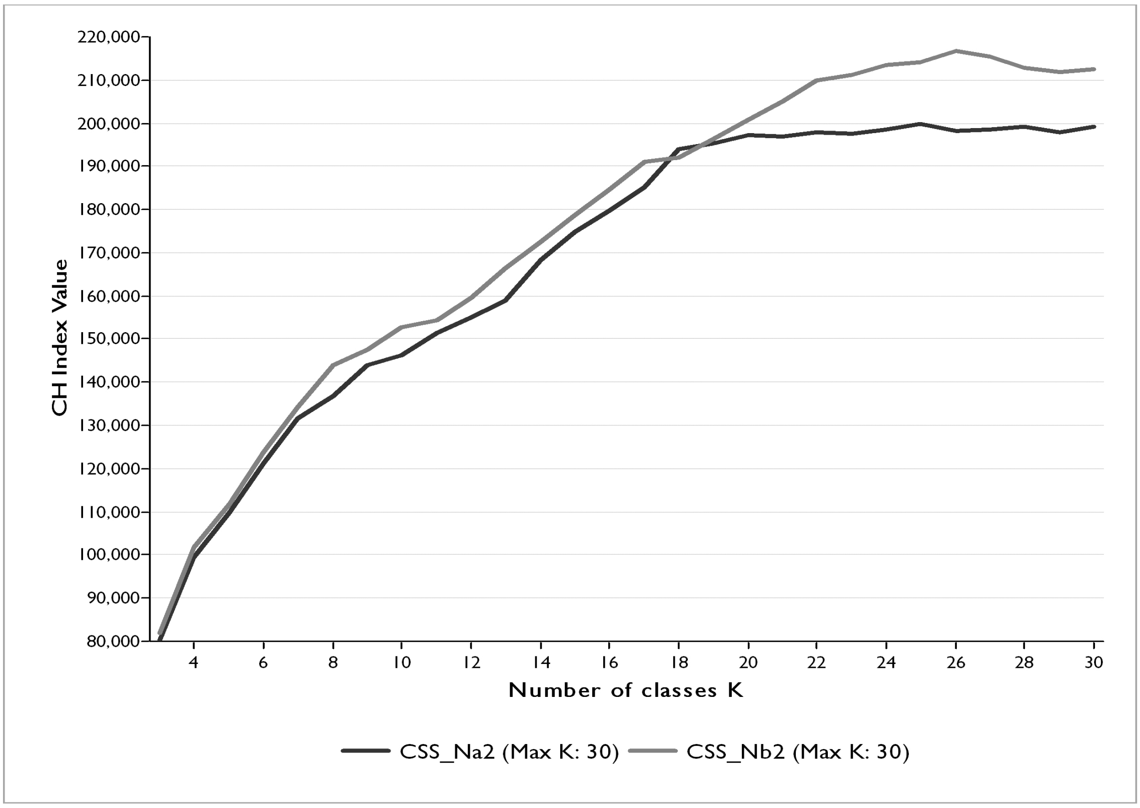


Figure 9: CH Index values for unsupervised classification of Combined Settlement Scores Na2 and Nb2

Table 17 lists the sets of clustered settlement variables and the 'best' number of clusters chosen for each set.

Table 17: Clustered settlement variables and 'best' numbers of clusters

Clustered Settlement Variables	'Best' Number of Clusters (K)
Distance to category ABCDE nucleations	5, 6
Distance to category BCD nucleations	4, 6
Combined Settlement Score Na2	5
Distance to category ABCDE nucleations combined with dispersion scores and hamlet counts	3, 10
Combined Settlement Score Nb2	5
Distance to category BCD nucleations combined with dispersion scores and hamlet counts	3

Maps of the clustered sets of assorted settlement variables using different values for K are presented in Figure 10–Figure 18. As would be expected, the higher the number of target clusters used by the ISODATA algorithm, the more spatially fragmented the resulting maps. That having been said, in all cases, the mapped clusters are considerably more compact and spatially coherent than those for the distance to all nucleations where K = 9 or for CSS Na2 and Nb2 where K = 25 or 26. The mapped clusters for the distance to nucleation categories A–E and B–D on their own and those for nucleation categories A–E combined with dispersion scores and hamlet counts using K = 10 (Figure 10–Figure 13 and Figure 17) exhibit a tendency toward 'bull's eyes' around some of the source nucleation points. All of the maps also show a certain degree of 'speckling,' where single grid cells or very small groups of cells of one value are surrounded by cells of another value. The maps of the clustered CSS variables (Figure 14 and Figure 15) are particularly prone to this effect. This 'speckling' could be smoothed away (see Lillesand *et al* 2008, 580-1, for methods) on the grounds that it is an artefact of the classification process. That may be true, but the unsmoothed maps highlight how the nature of historic settlement organisation could vary over fairly short distances. Reducing the amount of variability in the maps by smoothing away the 'speckling' would reduce the amount of information the maps contain, arguably hiding a genuine degree of local heterogeneity in settlement organisation. Future work could investigate the effects of smoothing or 'de-speckling' the maps of clusters.

What is clear from all the maps is that statistically similar clusters of rural settlement organisation were present in widely disparate locations across England. For example, when classifying nucleation categories A–E, the region of eastern Kent and the isle of Thanet is in the same statistical cluster as, among other places, southern Lancashire, much of the Nene valley, and an area along the north-east coastal plain stretching from Tyneside to Teesside and extending into south-western county Durham (cluster 2 in Figure 10 and Figure 11). This is not to suggest that the nature of rural settlement in such statistically similar but geographically distant places was identical, rather that, on the basis of the measures employed here, such areas were more like each other than like adjacent regions.

Figure 10–Figure 18 also show, for comparison, the clustered settlement variables overlaid with the outlines of Roberts and Wrathmell's settlement provinces, sub-provinces and local regions. In all the maps, there are some locations where the clustered settlement variables agree with Roberts and Wrathmell's boundaries but also numerous locations where they do not. The match between clusters for nucleation categories B–D where K = 4 or 6 and Roberts and Wrathmell's boundaries is particularly poor (Figure 12 and Figure 13). Of all the sets of clustered variables, the clustered Combined Settlement Scores and nucleation categories A–E and B–D combined with dispersion scores and hamlet counts (Figure 14–Figure 18) correspond most closely with the provinces, sub-provinces and local regions of the *Atlas*. Close inspection of the maps, however, reveals that even these cluster outlines and Roberts and Wrathmell's boundaries diverge more often than they agree.

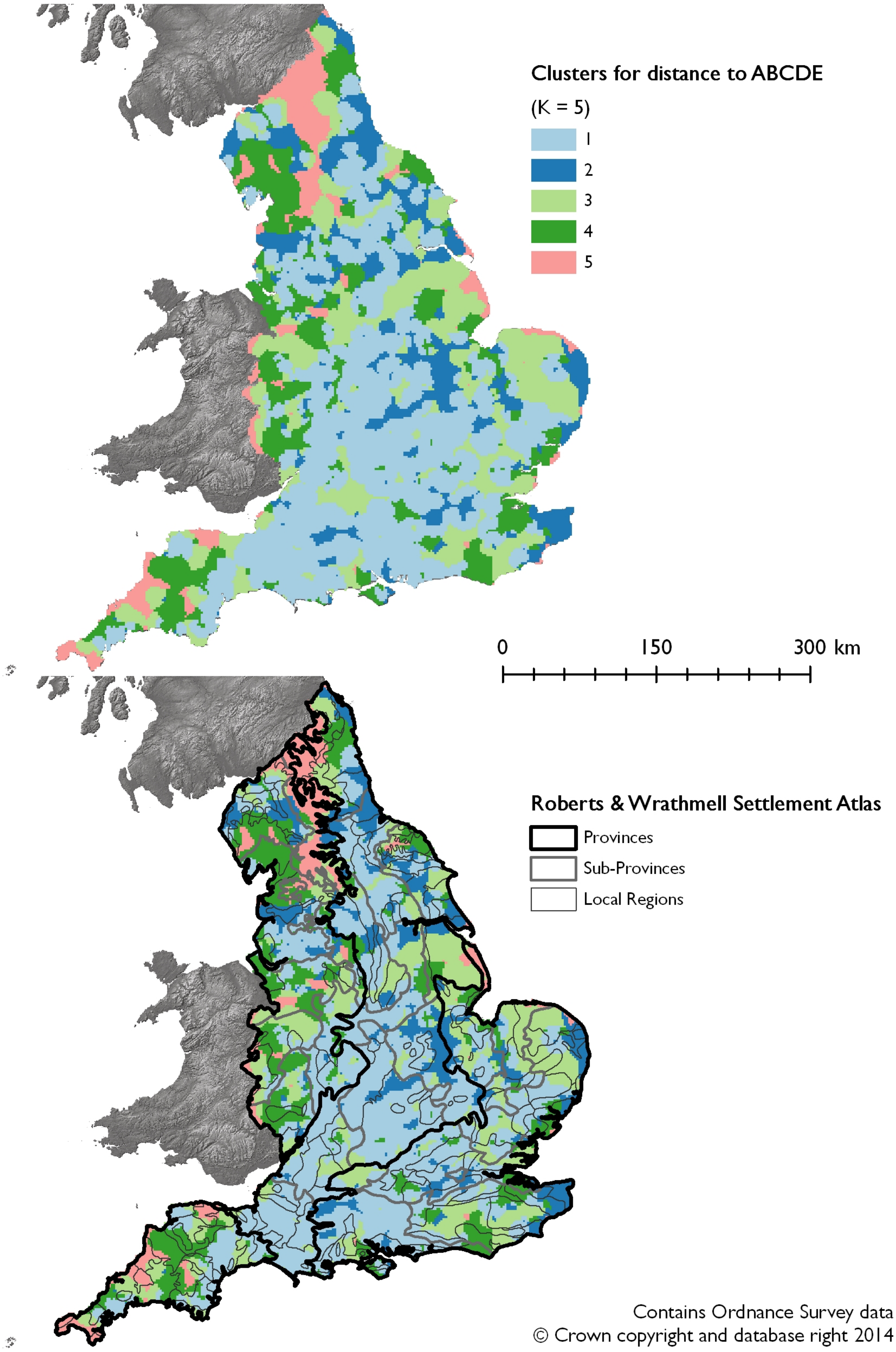


Figure 10: Maps of clusters for distance to category A–E nucleations using K = 5 (top) and the same overlaid with Settlement Atlas boundaries (bottom)

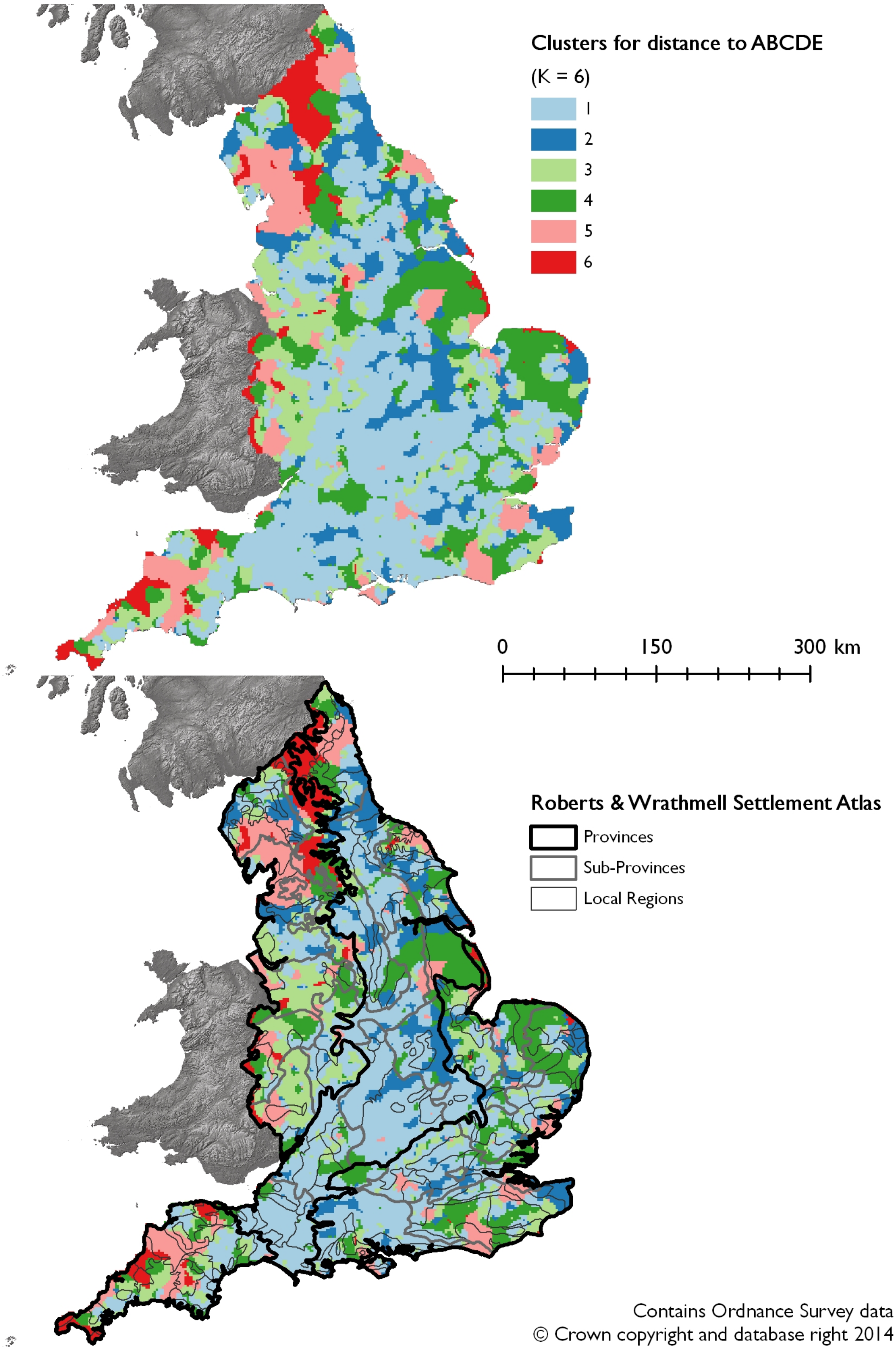


Figure 11: Maps of clusters for distance to category A–E nucleations using K = 6 (top) and the same overlaid with Settlement Atlas boundaries (bottom)

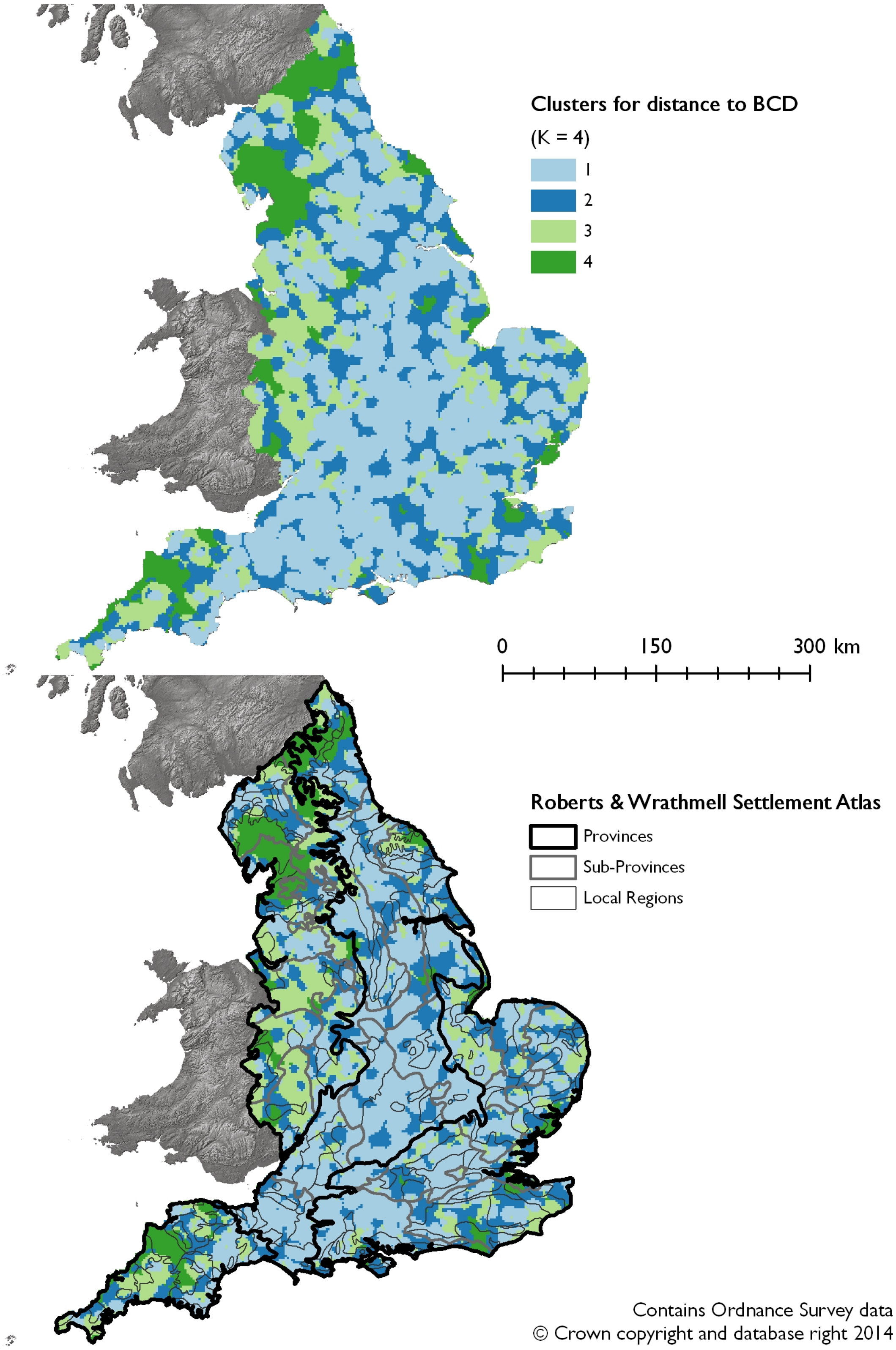


Figure 12: Maps of clusters for distance to category B–D nucleations using K = 4 (top) and the same overlaid with Settlement Atlas boundaries (bottom)

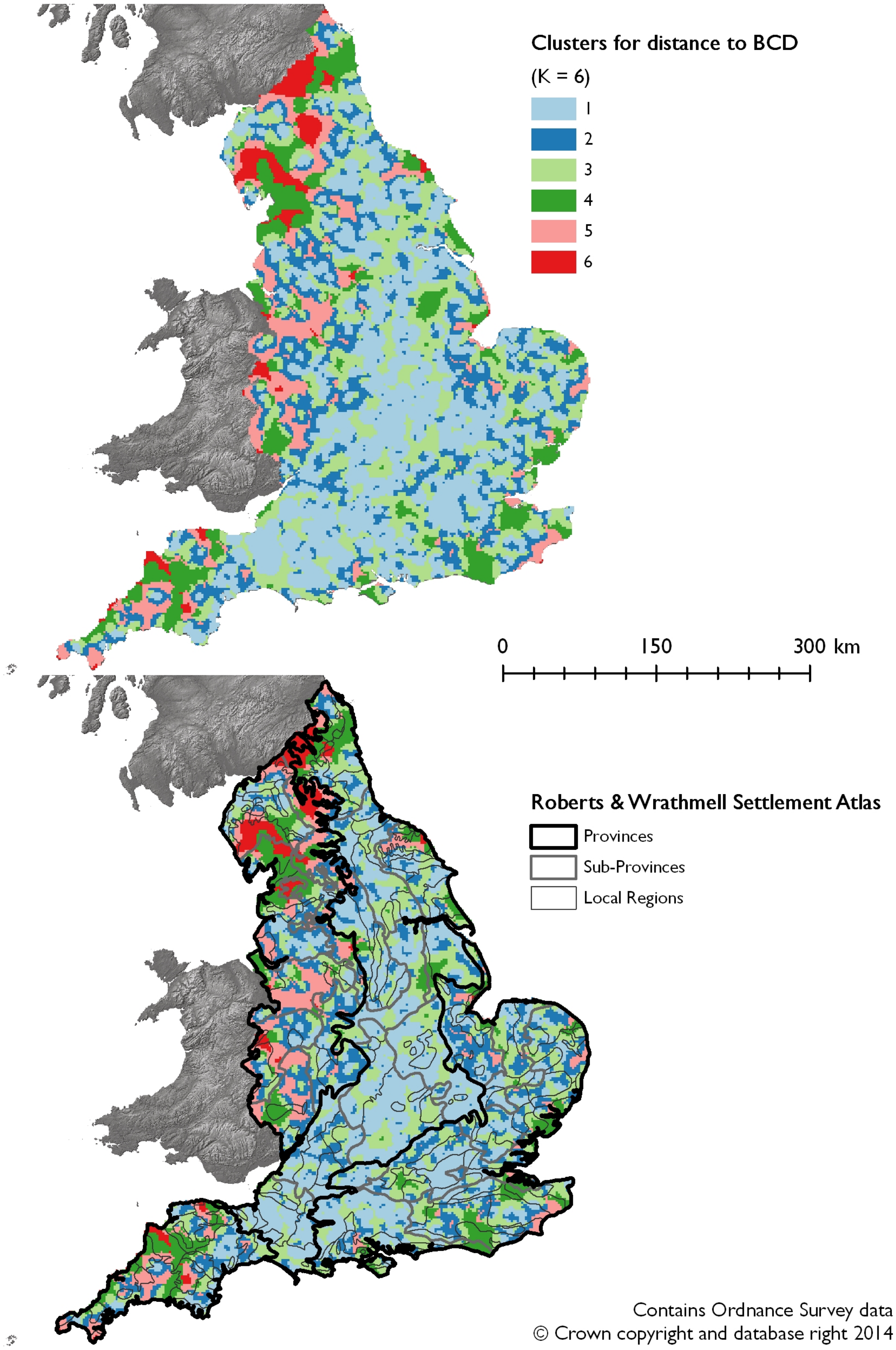


Figure 13: Maps of clusters for distance to category B–D nucleations using K = 6 (top) and the same overlaid with Settlement Atlas boundaries (bottom)

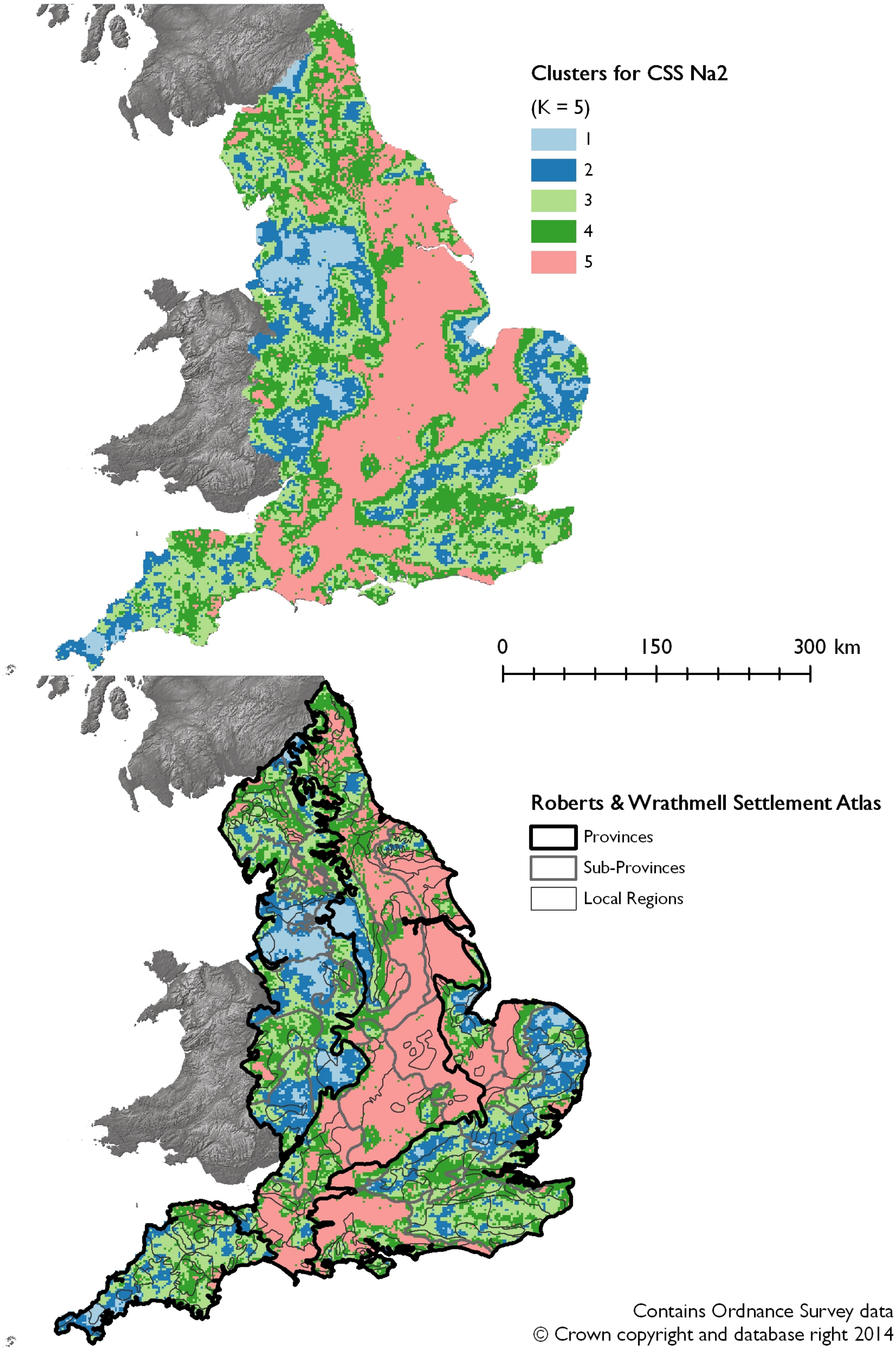


Figure 14: Maps of clusters for CSS Na2 using K = 5 (top) and the same overlaid with Settlement Atlas boundaries (bottom)

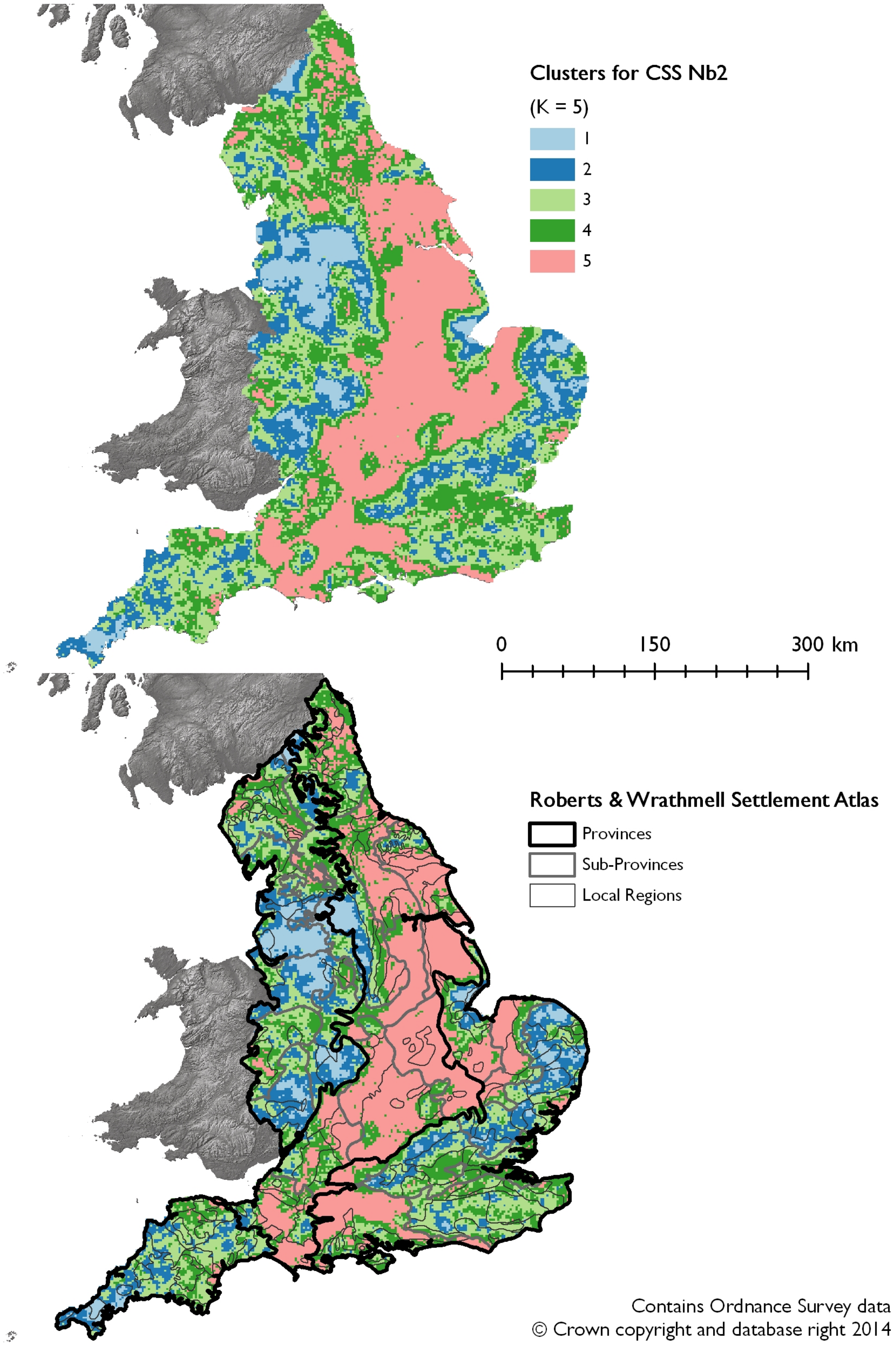


Figure 15: Maps of clusters for CSS Nb2 using K = 5 (top) and the same overlaid with Settlement Atlas boundaries (bottom)

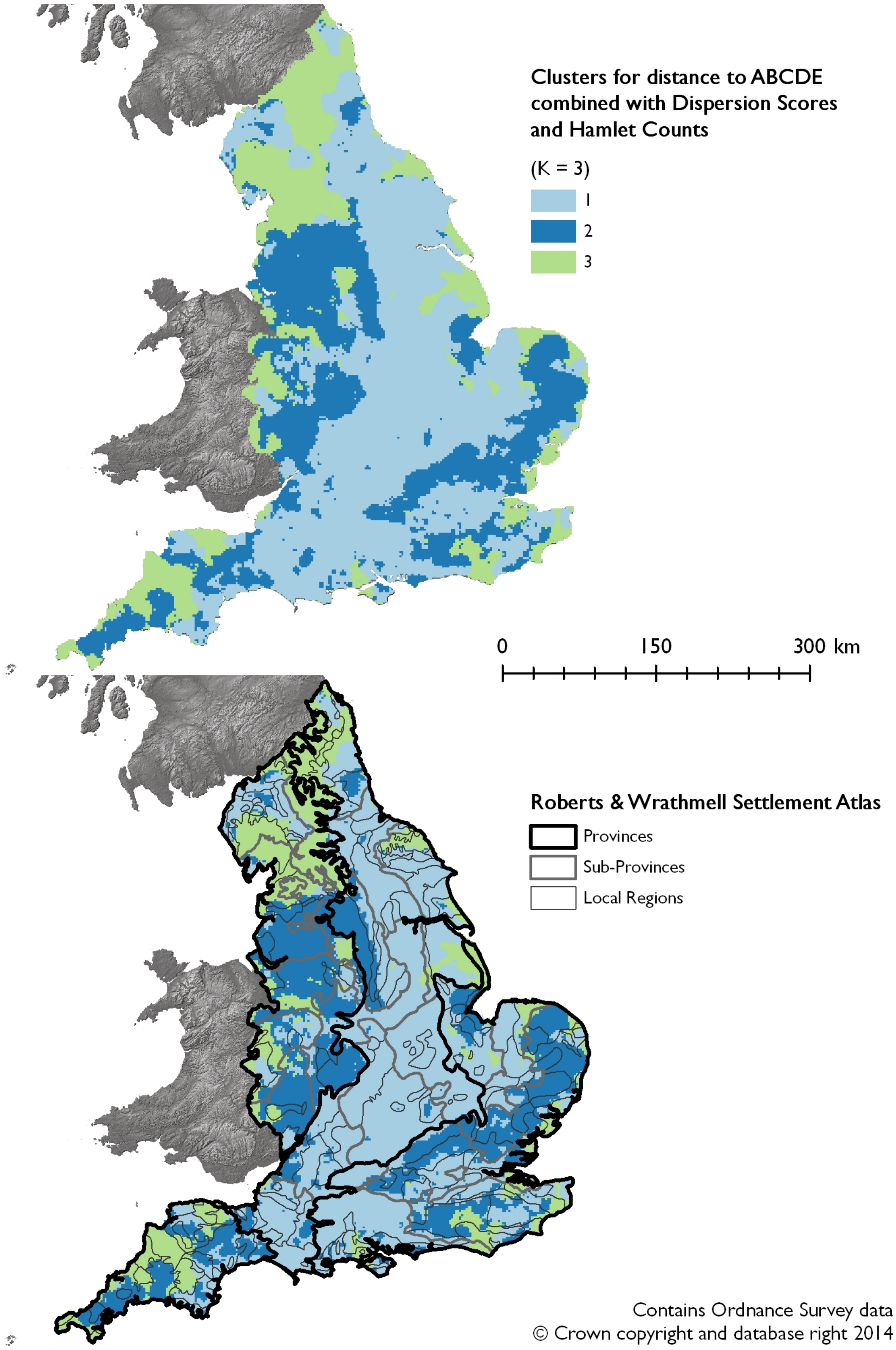


Figure 16: Maps of clusters for distance to category A–E nucleations combined with dispersion scores and hamlet counts using K = 3 (top) and the same overlaid with Settlement Atlas boundaries (bottom)

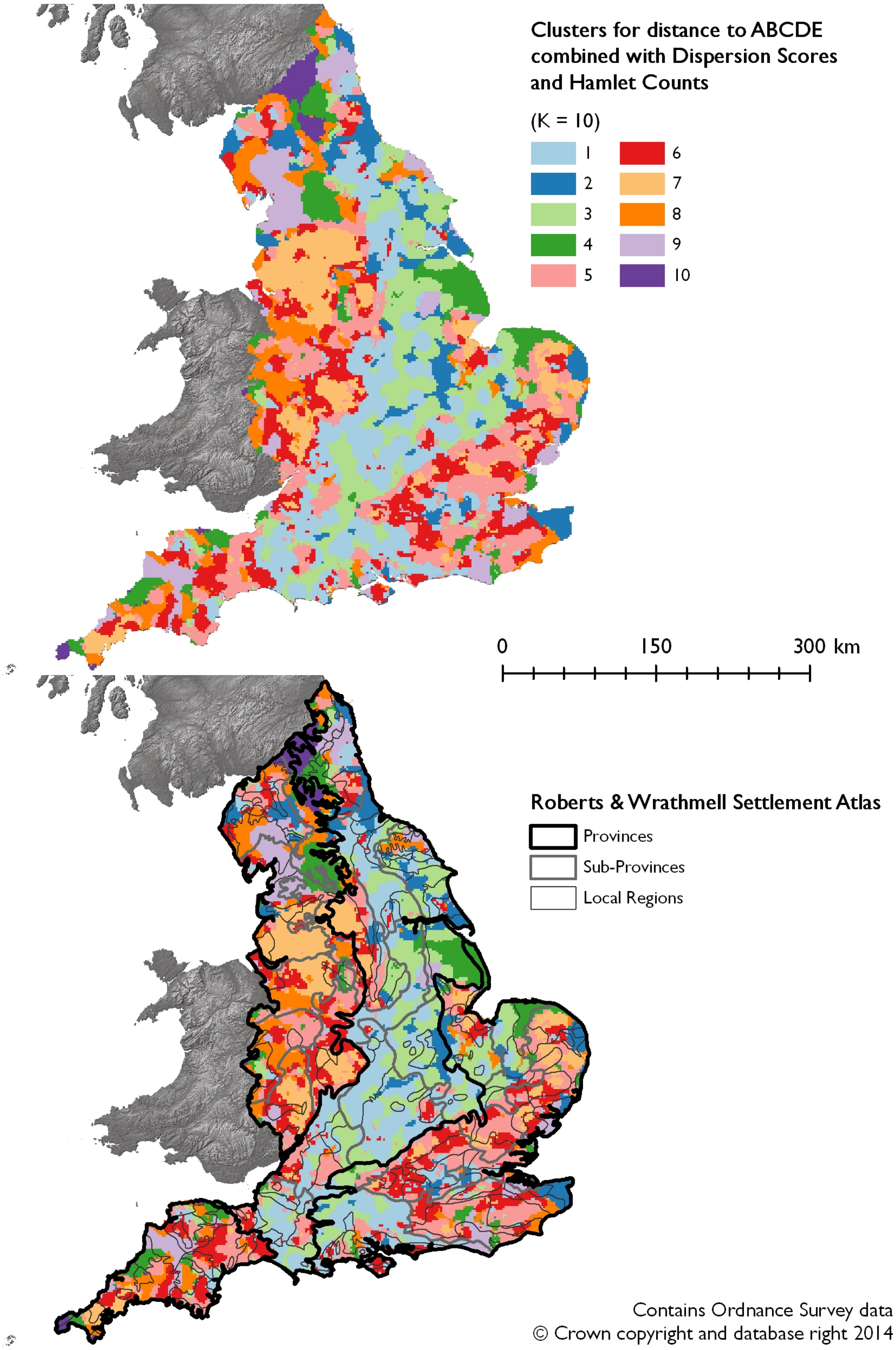


Figure 17: Maps of clusters for distance to category A–E nucleations combined with dispersion scores and hamlet counts using K = 10 (top) and the same overlaid with Settlement Atlas boundaries (bottom)

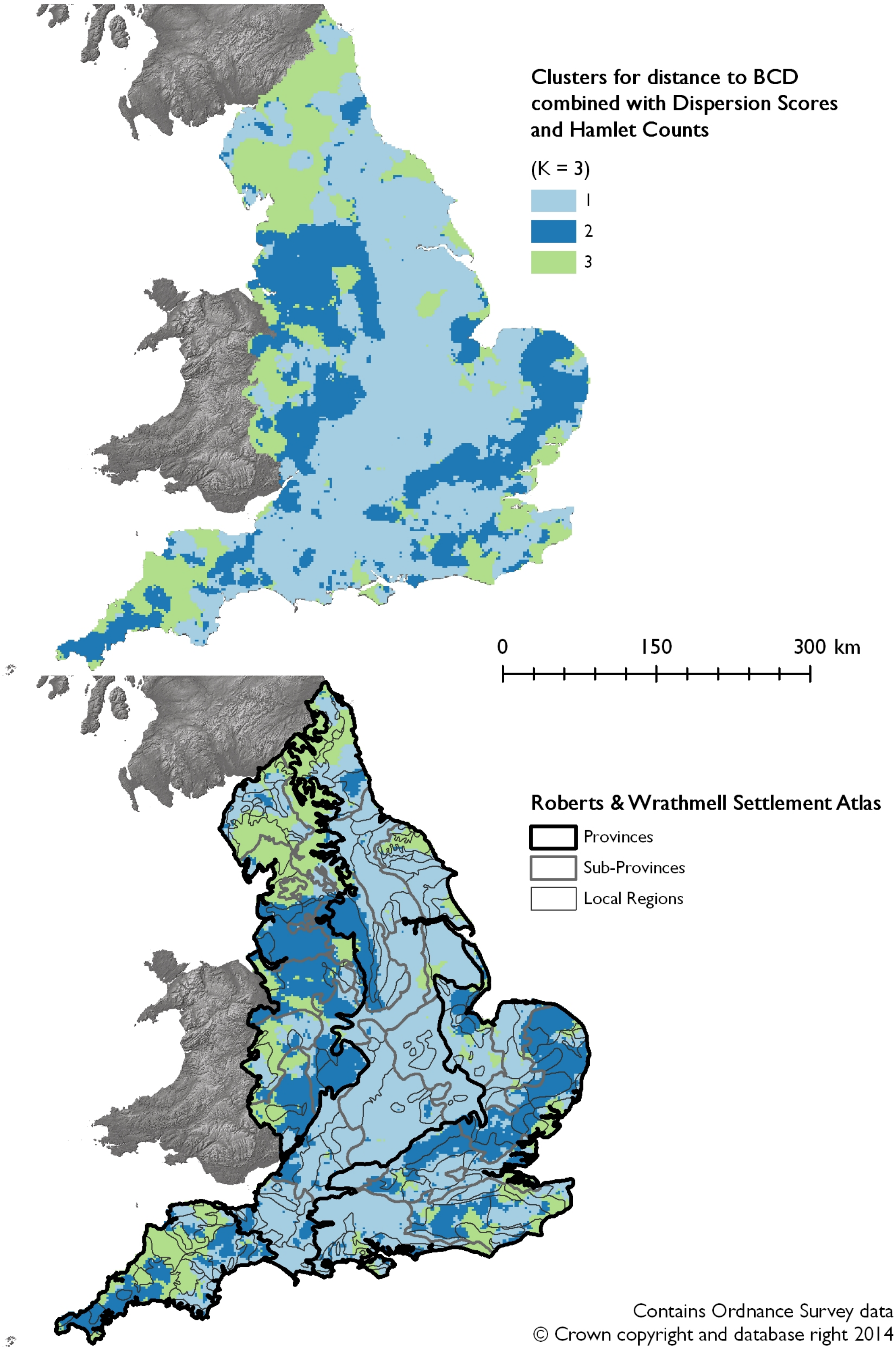


Figure 18: Maps of clusters for distance to category B–D nucleations combined with dispersion scores and hamlet counts using K = 3 (top) and the same overlaid with Settlement Atlas boundaries (bottom)

I also generated classified layers based on four sets of environmental variables used in the best-performing models in the regression analysis. Table 18 sets out the groups of variables used and the corresponding settlement variable.

Table 18: Sets of environmental variables used for unsupervised classification and 'best' numbers of clusters identified for each set

Set	Environmental Variables Used	Based on regression model for	'Best' Number of Clusters (K)
1	Elevation; p_av2m45; p_11; t_8; Soils Type 5; Soils Type 6; Soils Type 8; Soils Type 9; Soils Type 13; Soils Type 17; Soils Type 18; Soils Combo 32; Soils Combo 33; Soils Combo 35; Soils Combo 52; Soils Combo 53; Soils Combo 54	DstNclAll	5, 8
2	Elevation; p_av2m34; p_11; DSR_av2m78; Soils Type 3; Soils Type 6; Soils Type 8; Soils Type 9; Soils Type 13; Soils Type 17; Soils Type 18; Soils Combo 32; Soils Combo 33; Soils Combo 35; Soils Combo 52; Soils Combo 53; Soils Combo 57	DstNclBCD	5, 7
3	Elevation; p_av3m345; p_11; t_8; Soils Type 3; Soils Type 6; Soils Type 8; Soils Type 9; Soils Type 13; Soils Type 17; Soils Type 18; Soils Combo 32; Soils Combo 33; Soils Combo 35; Soils Combo 52; Soils Combo 53; Soils Combo 57	DstNclBCD	5, 11
4	Elevation; p_3; p_9; bio1; Soils Type 3; Soils Type 5; Soils Type 6; Soils Type 7; Soils Type 8; Soils Type 9; Soils Type 13; Soils Type 17; Soils Type 18; Soils Combo 32; Soils Combo 33; Soils Combo 35; Soils Combo 52; Soils Combo 53	CSS Na2, CSS Nb2	3, 9

Figure 19 illustrates the CH Index values for different values of K used when classifying the sets of environmental variables. Again, arrows point out the CH Index values indicating the values of K which I took to indicate the best numbers of clusters, which are also noted in Table 18. For all of the sets of environmental variables, more than one value of K appeared reasonable.

For comparison, Figure 20 shows Roberts and Wrathmell's detailed terrain zones, which characterise the landscape based on 'relief, drainage, country rock and drift, mountain peaks and escarpments, plateaux and ridges, lowland and plain, and marsh and fen' (Roberts and Wrathmell 2000, 16). Maps of the clustered sets of assorted environmental variables using different values for K are presented in Figure 21–Figure 24. Again, higher numbers of target clusters used by the ISODATA algorithm produced more spatially fragmented maps. The same issues of 'speckling' in the maps made above regarding the clustered settlement variables also apply here. As with the settlement clusters, I deliberately chose not to smooth the maps of clustered environmental variables on the grounds that doing so would result in a loss of information. Unsurprisingly, given the use of quite similar sets of input variables, some of the different maps of clustered variables look much alike. Comparing Roberts and Wrathmell's terrain zones with the maps of clustered environmental variables also reveals many broad similarities, but the correspondences are inexact.

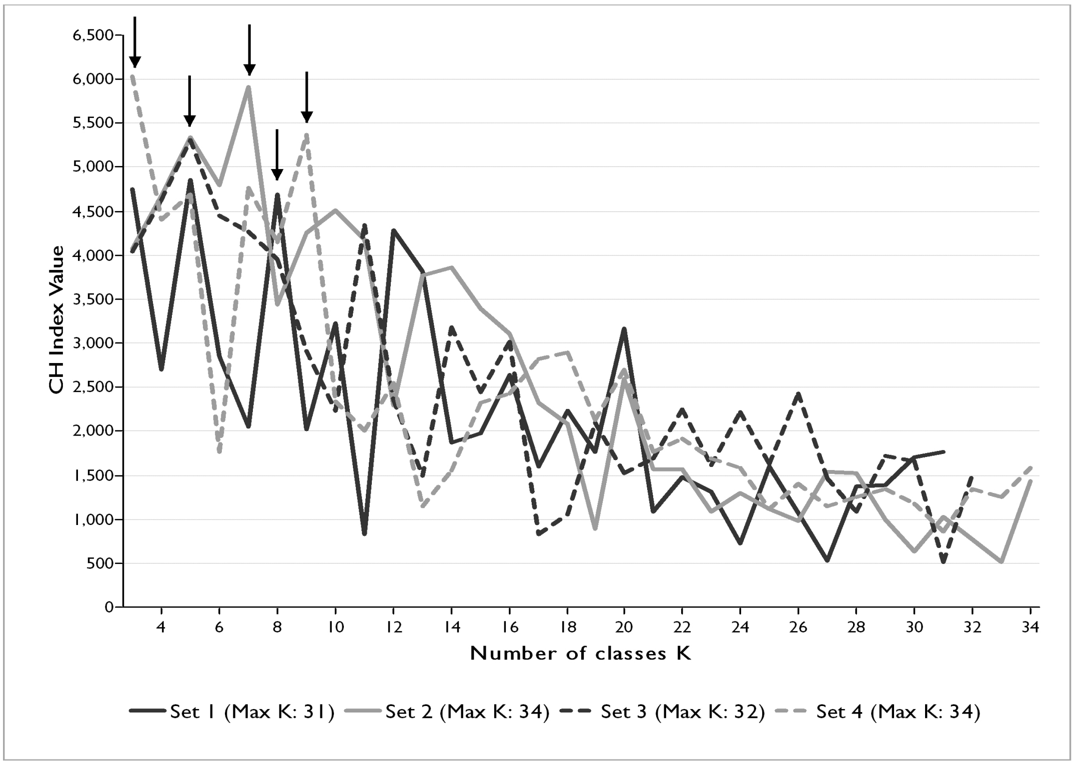


Figure 19: CH Index values for unsupervised classification of environmental variable sets

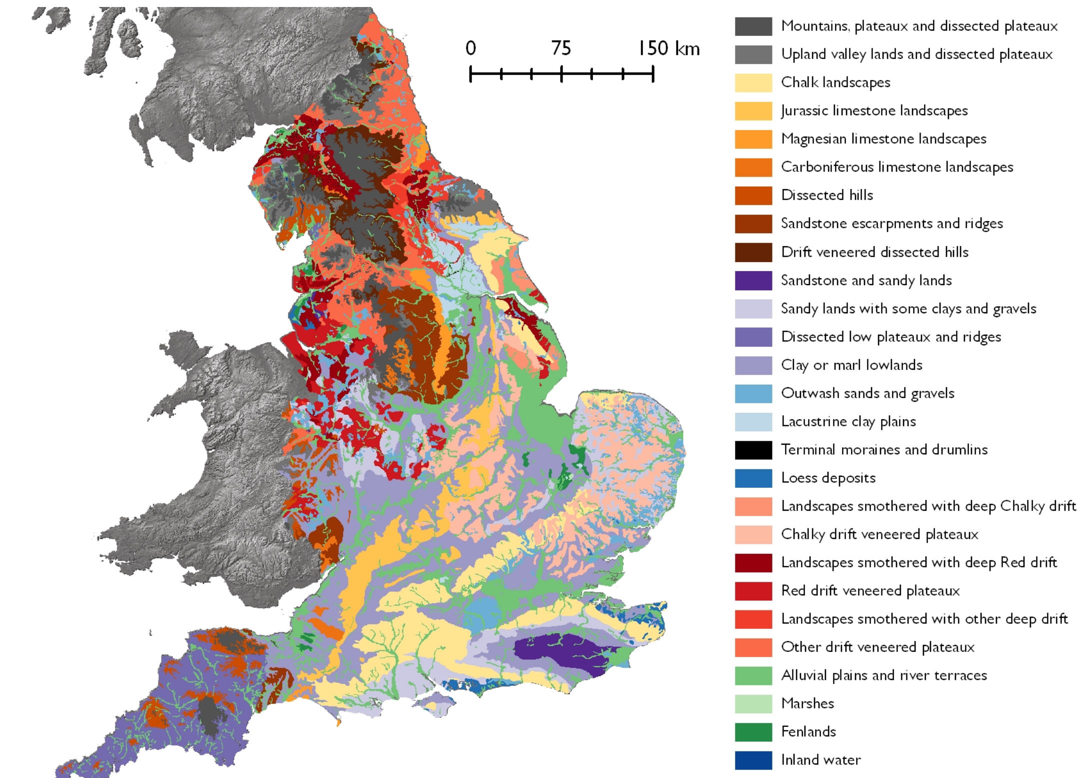


Figure 20: Terrain Zones from Roberts and Wrathmell's Atlas

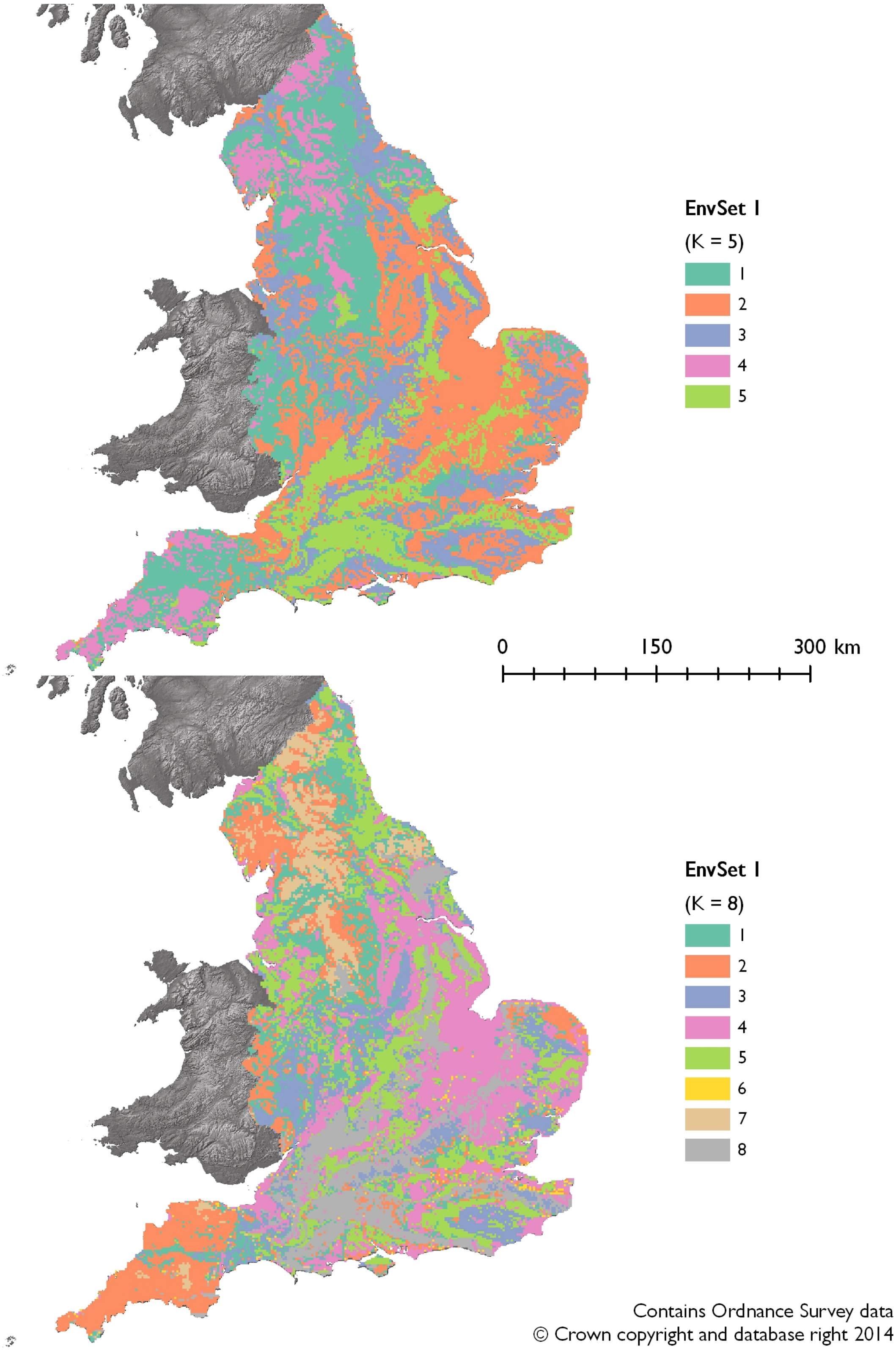


Figure 21: Maps of clusters for environmental variable Set 1 using K = 5 (top) and K = 8 (bottom)

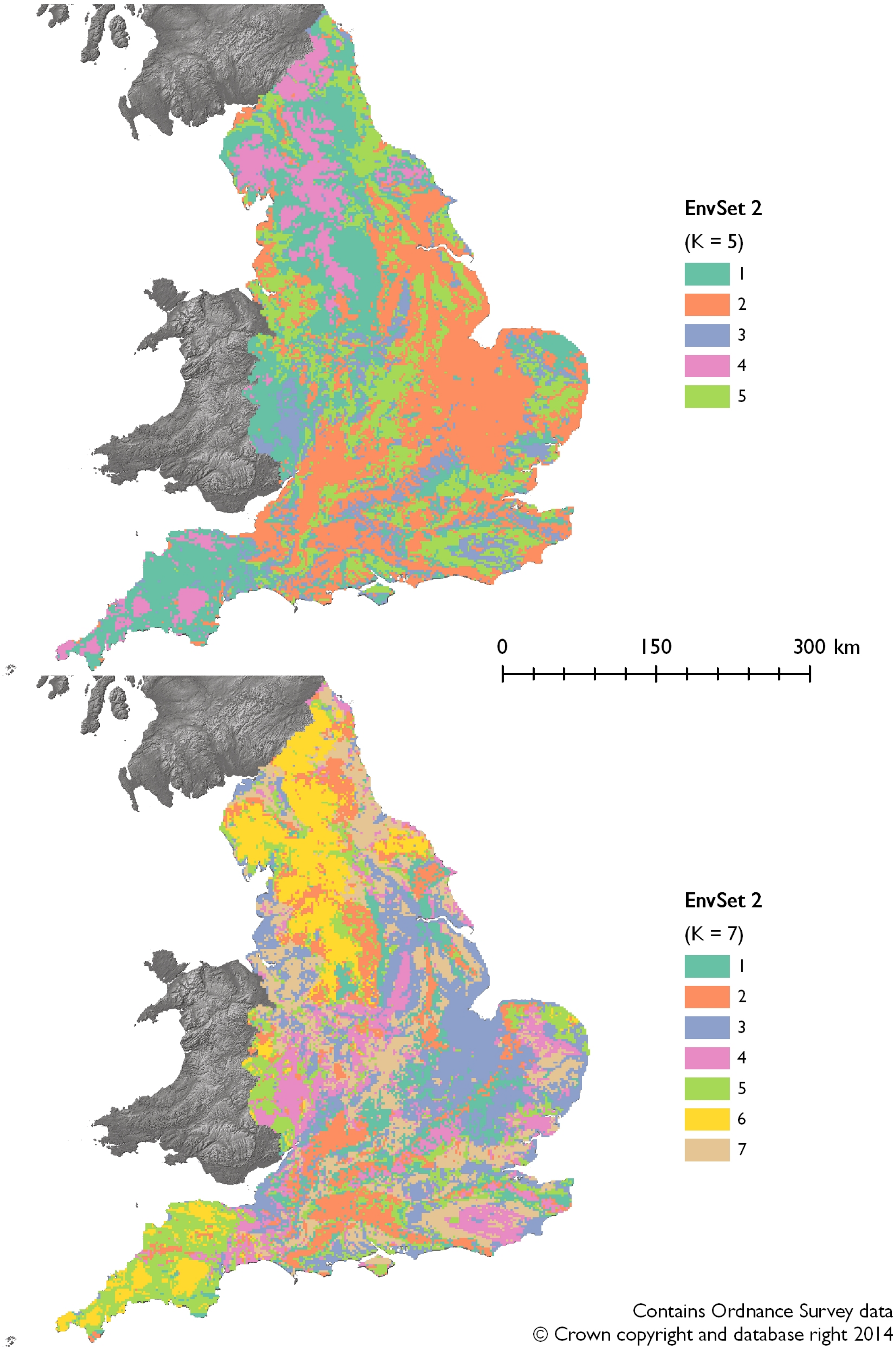


Figure 22: Maps of clusters for environmental variable Set 2 using K = 5 (top) and K = 7 (bottom)

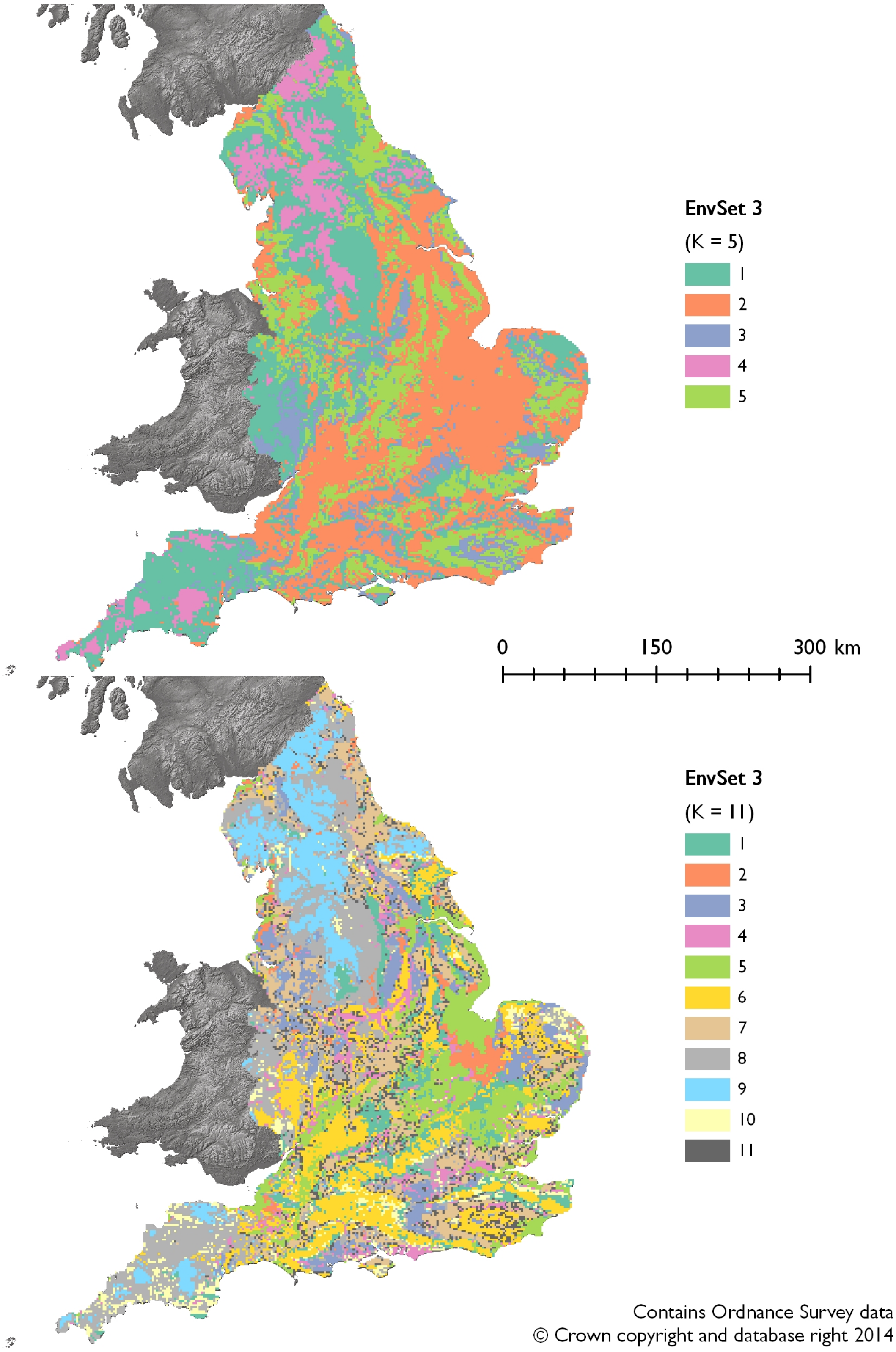


Figure 23: Maps of clusters for environmental variable Set 3 using K = 5 (top) and K = 11 (bottom)

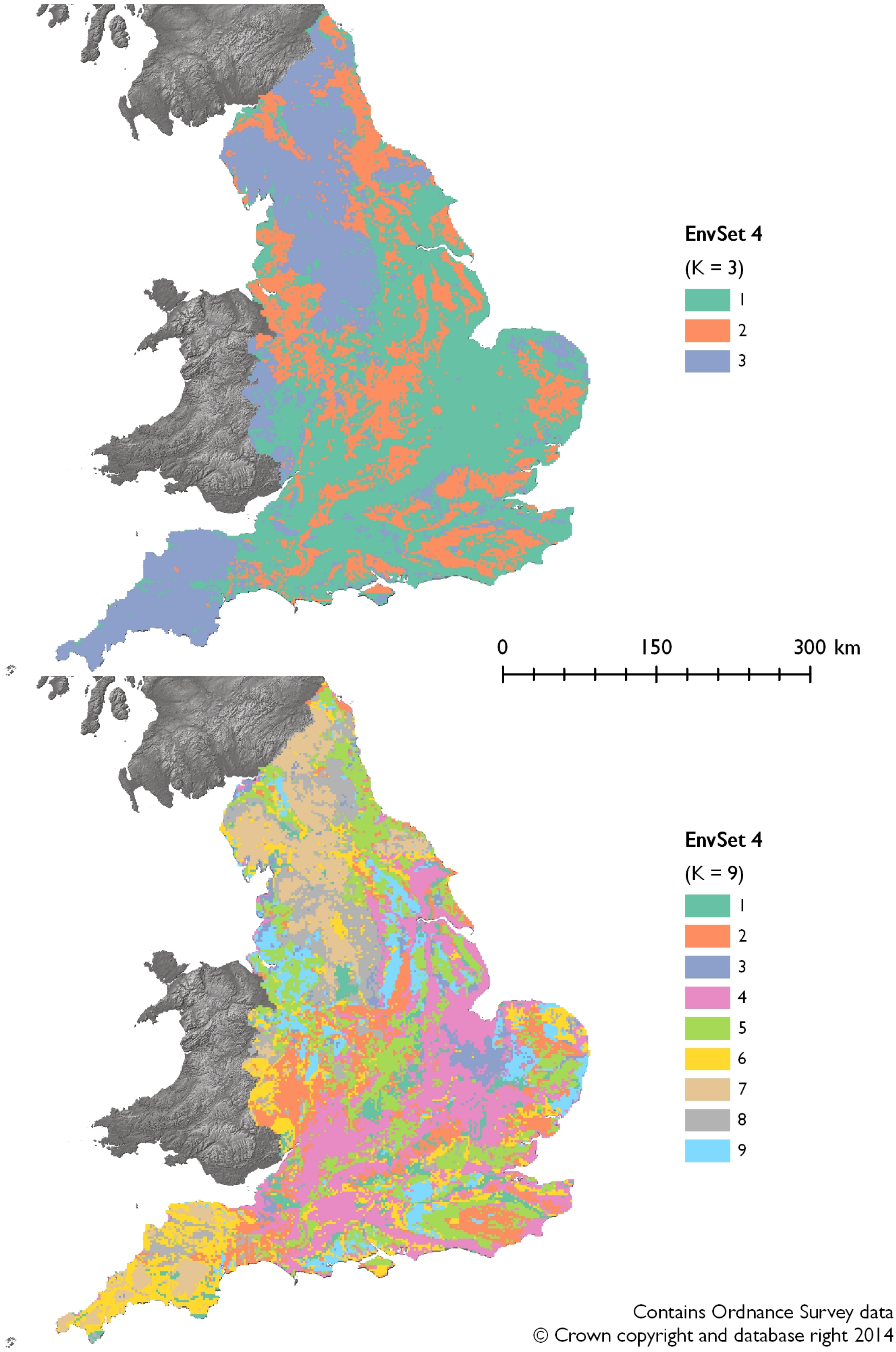


Figure 24: Maps of clusters for environmental variable Set 4 using K = 3 (top) and K = 9 (bottom)

Discussion

The maps of neither the clustered settlement variables nor the clustered environmental variables correspond closely with the boundaries of settlement regions and terrain zones delineated by Roberts and Wrathmell.

The lack of precise agreement between the clustered environmental variables and Roberts and Wrathmell's terrain zones is perhaps unsurprising, given the differing sources and methods. They defined their zones primarily on the basis of bedrock and superficial geology, rather than soils, and they did not consider temperature and precipitation at all. Of course, bedrock and superficial geology are major determinants of soil composition – they are so-called 'parent materials' (Lawley 2009) – so some resemblance between the maps would be expected. Roberts and Wrathmell's sources and those used here, are, however, different enough to explain much of the differences in results. Additionally, Roberts and Wrathmell depicted 27 different terrain categories or 25 if one disregards inland water and terminal moraines and drumlins as not relevant or too localised to be meaningful when viewed at a national scale. The number of clusters used for the unsupervised classification ranged from 3 to 11, meaning that the resulting maps are inevitably simpler than Roberts and Wrathmell's. Merging some of their zones might produce maps more closely resembling the maps of clustered environmental variables. Roberts and Wrathmell's choices of what and how many categories of terrain to map were the product of much research and judgement (Roberts and Wrathmell 2000, 16-17), but as with the settlement maps, their process was one of 'little science but much logic' (ibid, 13). The environmental variables I used were chosen based on the detailed model selection procedure described in the sections above. I do not claim that my results are inherently better, but using the data and methods described here, my results would be directly reproducible, which cannot be said of Roberts and Wrathmell's maps.

The mismatches between the maps of clustered settlement variables and Roberts and Wrathmell's provinces, sub-provinces and local regions are arguably more noteworthy. The differences in results may simply be the products of the differing methods used to create the regions. The sources, however, are the same. Roberts and Wrathmell freely admitted subjectivity in drawing their boundaries, and allowed that a repeat of the mapping process they used would likely produce somewhat different results (ibid, 16). My results clearly demonstrate that one can group the source data in multiple – and arguably equally legitimate – ways. Again, I do not claim that my results are necessarily better than Roberts and Wrathmell's. Future work could, however, quantify the performance of Roberts and Wrathmell's regionalisation of the settlement data using the CH Index (or some other goodness-of-fit measure) to assess whether their division of the landscape is statistically better or worse than the clustered maps produced here. And as with the clustered environmental variables, my results would be directly reproducible using the data and methods set out here.

Having clustered and mapped the sets of settlement and environmental variables, it becomes possible to explore more directly the distribution of the former in relation to the latter. For example, Figure 25 shows, on the top, clusters for distance to category A–E nucleations combined with dispersion scores and hamlet counts using K = 10 and, on the bottom, those for environmental variable Set 4 using K = 9. Visual comparison of the two maps suggests the pattern of boundaries in the clustered environmental data is broadly similar to that in the clustered settlement data. The distribution of the settlement clusters labelled 1, 2 and 3 across central England roughly follows the distribution of environmental clusters 3 and 4. Closer inspection, however, reveals that the pattern of environmental factors clusters is not strongly replicated in the settlement clusters. Comparison is challenging, however, even when the maps of the sets of clusters are immediately adjacent on the page. Figure 26 shows the same clustered environmental variables, this time overlaid with the boundaries of the settlement clusters shown in the top map in Figure 25. With the two sets of clusters depicted on a single map, it is possible to pick out areas where the boundaries of the polygons match up, but in a large number of locations, they do not.

Purely visual assessment of the relationships between the boundaries remains difficult, even having directly overlaid the boundaries of clusters derived from the settlement variables on those derived from the environmental variables. Moving beyond a fairly impressionistic interpretation of how the two sets of clusters match up requires setting out explicit methods and criteria to define the degree of match. To what extent does a polygon in one set of clusters need to overlap one in the other set for them to be said to 'agree'? On that basis (whatever it is), how many polygons do agree? How many disagree, ie, do not match up? If some polygons do agree and others do not, how much agreement is needed to enable one to conclude there was a meaningful degree of match? Perhaps most importantly, how likely it is that any agreement (or disagreement) between polygons might be the result of random chance? These questions cannot be answered in a robust fashion simply by 'eyeballing' the maps. In the next section, I discuss and present the results of using one possible method for investigating in a formal, comprehensive fashion the relationships between the clusters derived from the settlement variables and those derived from the environmental variables.

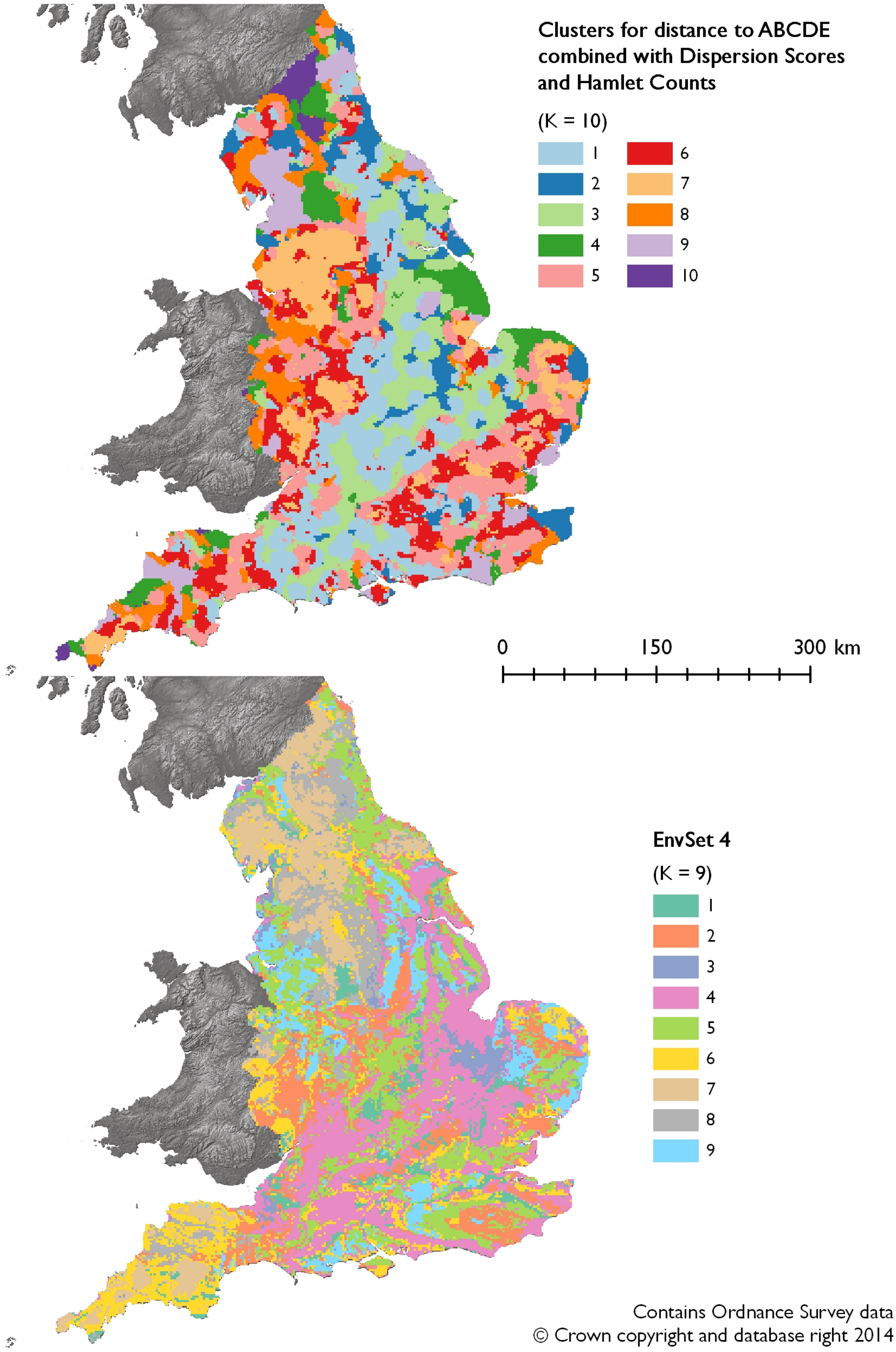


Figure 25: Maps of clusters for distance to category A–E nucleations combined with dispersion scores and hamlet counts using K = 10 (top) and environmental variable Set 4 using K = 9 (bottom)

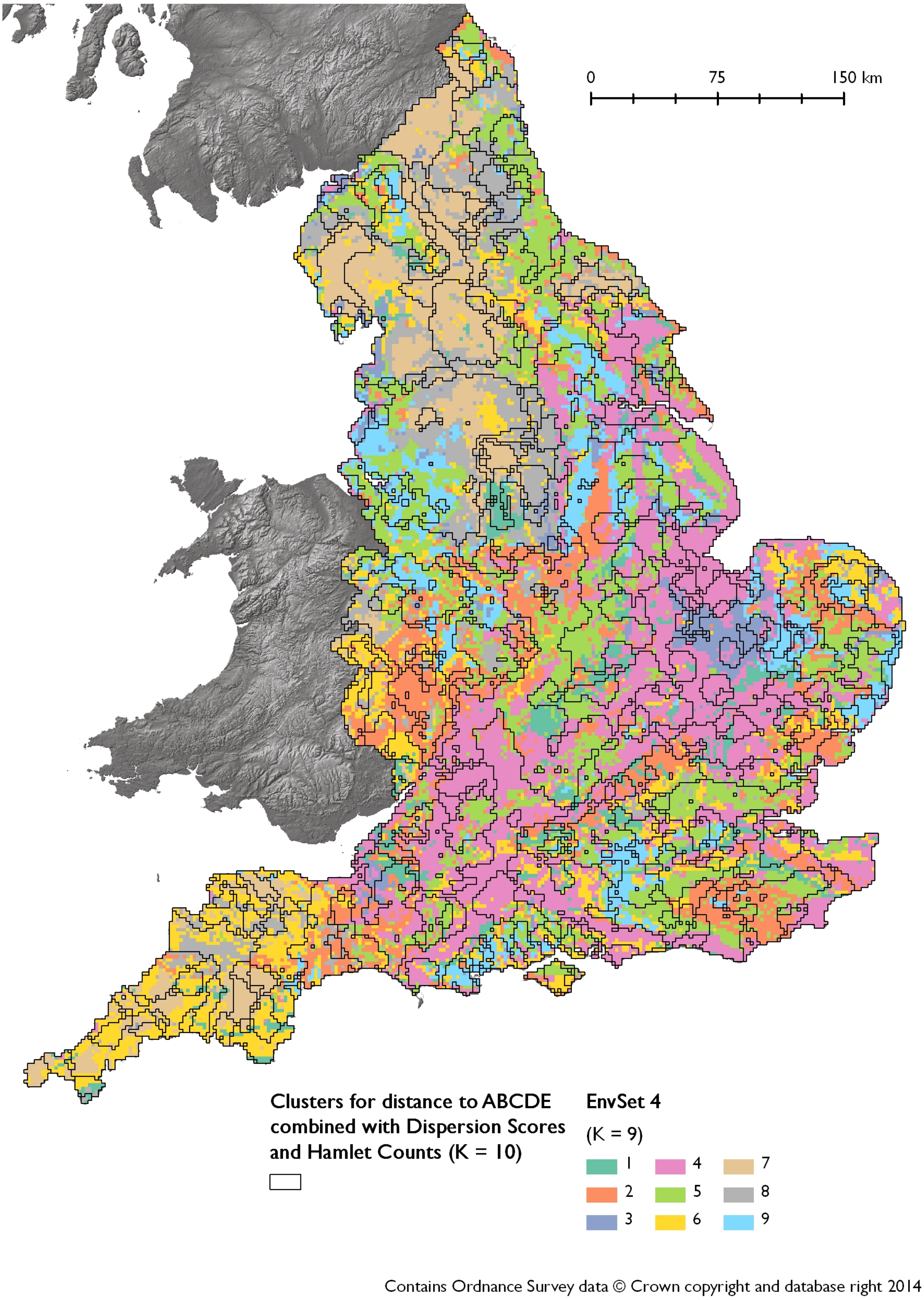


Figure 26: Map of clusters for distance to category A–E nucleations combined with dispersion scores and hamlet counts using K = 10 overlaid on clusters for environmental variable Set 4 using K = 9

Relative Area Overlap Analysis

As discussed earlier, regression analysis makes a number of assumptions about the numeric distributions of and the form of relationships between the variables, assumptions that the data used here do not necessarily meet. And as noted in the preceding section, simple, visual inspection of pairs of maps, or even single maps with the relevant features overlaid on one another, may not be sufficient to detect and reliably assess patterns and whether they match. There are, however, tests of spatial association that do not make rigid assumptions about the data and do explicitly look for correspondence or coincidence between the shapes and locations of features in sets of spatial data. The approach I employ here is a form of map comparison analysis (Haining 2003, 265-270), using a technique developed by Maruca and Jacquez (2002) for area-based tests of association between spatial patterns. As far as I know, this technique has not been incorporated into any publicly-available GIS or spatial analytical software. Where it has been used, investigators have implemented it using the statistical programming language R (eg, Sirami *et al* 2009; Fortin *et al* 2005), rather than ArcGIS, Python and ArcPy as done here. The technique makes use of the clustered or regionalised settlement data and environmental datasets discussed in the previous section, comparing the degree of overlap between them.

Method

The method aims to quantify how well two sets of polygons match, based on the Relative Area Overlap (RAO for short) between the individual polygons in each of the two partitions. Maruca and Jacquez test the significance of the amount of overlap using a Monte Carlo randomisation procedure. The method set out in the original paper provides a *global* RAO statistic, that is, one summarising RAO over the whole of the set of polygons. I have extended the method to calculate a *local* RAO statistic: one for each polygon in the set (Haining 2003, 186-7; Boots and Okabe 2007).

The method uses two sets of polygons as inputs, referred to as Set I and Set J, both of which are partitions of the same geographic space. Here the assumption is that Set J represents some phenomenon whose delineation is thought to have influenced the delineation of some other phenomenon, represented by Set I. For each pair of overlapping polygons in I and J, one calculates the ratio of the area of intersection of the polygons to the area of union of the polygons. Figure 27 shows an example of pairs of overlapping polygons and their individual RAO values. The coloured polygons labelled A and B comprise one set, eg, Set J, and the hatched polygons labelled 16 and 22 comprise the other, eg, Set I.

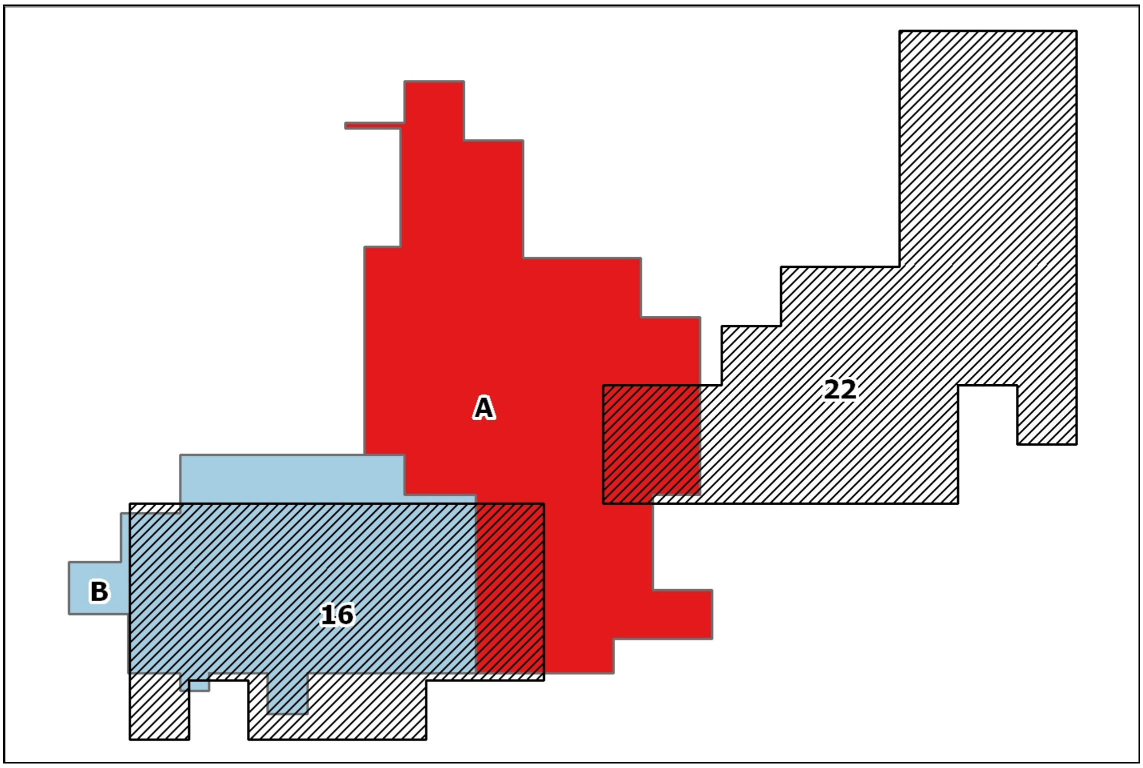


Figure 27: An example of Relative Area Overlap for pairs of polygons.

Overlap A : 16 = 0.05; A : 22 = 0.04; B : 16 = 0.37; B : 22 = 0.00

Each polygon in Set I may overlap with more than one polygon in Set J, so for each polygon in Set I, one finds the highest value for that ratio: the maximum relative area overlap. The global RAO for Set I is the average of the maximum relative area overlap for all the polygons in Set I when compared to Set J. One can also calculate global RAO statistics comparing how well the polygons in Set J overlap with those in Set I, as well as for the bi-directional overlap between the two sets.

To account for the possibility that the two partitions may have polygons of wildly different sizes, Maruca and Jacquez advocate calculating an area-weighted RAO statistic as well. Here, the weighting factor is the area of each polygon in Set I. It may be the case that many small polygons in Set I overlap closely with polygons in Set J, but many large polygons in Set I do not, suggesting that, overall, the correspondence between Sets I and J is arguably fairly low. The unweighted RAO statistic, however, may not adequately reflect this situation and could be misleading. The weighted RAO statistic may be a more reliable indicator of association in such cases.

Having calculated the global RAO statistics comparing Sets I and J, the problem then becomes evaluating how likely it is that the degree of overlap seen is simply the product of random chance, that is, test the statistical significance of the result. This is done through Monte Carlo randomisation. First, one creates a set of alternative, randomly generated partitions with the same number of polygons as Set I and occupying the same geographic space. Then one calculates the RAO statistic for each of the random partitions to generate a reference distribution. Large values for the RAO statistic indicate a high degree of overlap between the polygons in Set I and Set J, which suggests that the pattern of polygons in the two sets are associated. Small values for the statistic indicate little overlap, suggesting a lack of association. The results are ranked in ascending order, together with those from the original data, to obtain p-values. The p-value of each result is its rank divided by the total number of iterations. One can then compare the p-values to a particular level of statistical significance (α), say 0.05 or 0.10, to decide whether or not the result is unusual enough to reject the null hypothesis that there is no association between the sets of polygons.

Maruca and Jacquez use randomly generated Voronoi (Thiessen) polygons for their alternative partitions, which is perfectly reasonable and similar to a number of other spatial statistical techniques. There are, however, two notable issues with this approach, both of which Maruca and Jacquez touch on in their paper. First, creating the alternative partitions in this way is predicated on a null hypothesis of complete spatial randomness (CSR), which is arguably not appropriate in this case. It is abundantly clear that the settlement data are not randomly distributed across the landscape, that is, they are positively spatially autocorrelated. To test data known and expected to be spatially autocorrelated against a null hypothesis of CSR is nonsense. The reference distribution must be generated from randomisations that reasonably reflect the processes that generated the actual data (Waller and Jacquez 1995; Fortin and Jacquez 2000). Second, Voronoi polygons, while widely used in a variety of applications (Okabe *et al* 2000), often do not adequately reflect the shapes of polygons mapping real spatial phenomena, so they can be a poor basis for comparison (see, eg, Gregory and Ell 2007, 68-70).

To address these issues, and simultaneously enable the calculation of local RAO statistics and p-values, I have taken a different approach to creating the alternative polygon partitions. I began by spatially 'shuffling' the polygons in Set I, randomly moving each polygon to the centroid of another polygon. I then cleaned and flattened the shuffled polygons to match the footprint of the original set, eliminated 'sliver' polygons, and then copied the unique IDs from the original set onto the cleaned and flattened 'shuffled' polygons. The process is akin to taking the pieces of a jigsaw puzzle and randomly squashing them into place until the extent of the puzzle is filled, but paying no attention to whether the pieces actually fit together. Figure 28 illustrates a simulated source set of polygons on the left and one example of a randomly shuffled, cleaned and flattened set on the right. Close comparison of the two sets can reveal which shuffled polygon is derived from which source polygon. For example, shuffled polygon 3 is derived from source polygon 12, shuffled polygon 28 from source polygon 23, and shuffled polygon 26 from source polygon 20. Of course, the shapes of many of the shuffled polygons are considerably different from those in the source set. Overall, however, the similarity between the shuffled and source polygons is considerably greater than would be the case if comparing completely randomised Voronoi polygons and the source set.

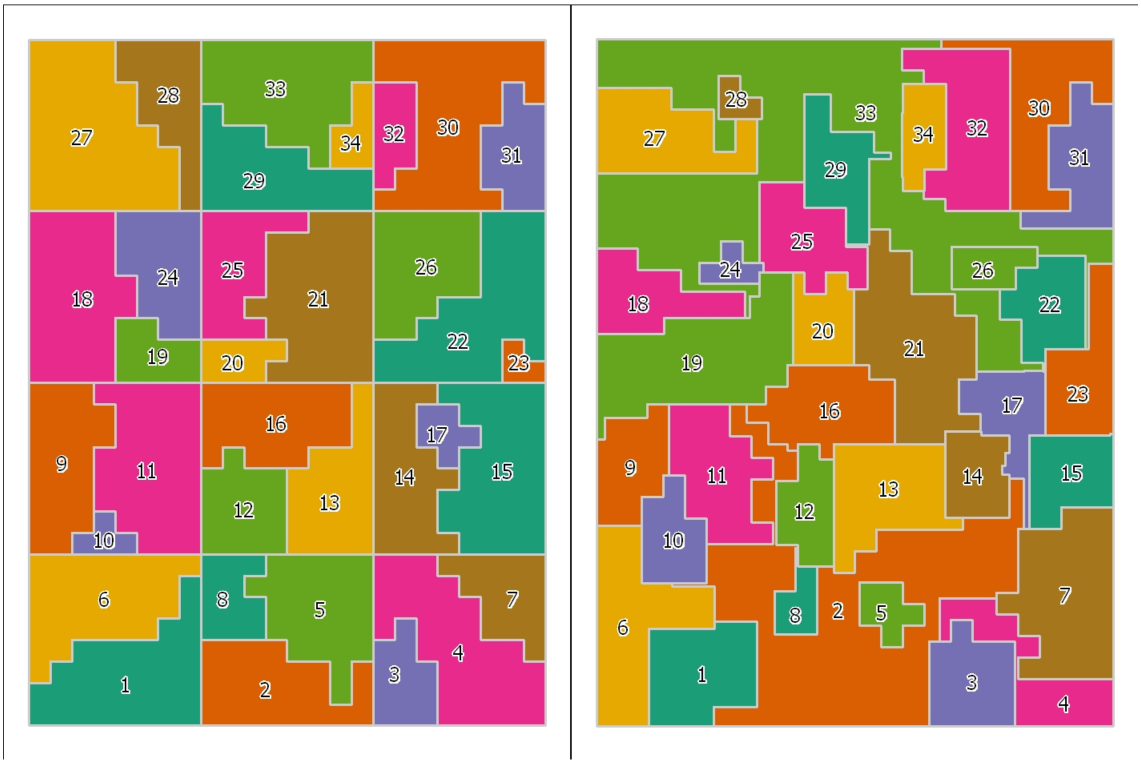


Figure 28: A simulated source set of polygons (left), and an example randomly shuffled, cleaned and flattened set (right)

Re-using the source polygons means the randomised shapes more realistically reflect the varying sizes and shapes of the real data than would Voronoi polygons. Focusing each shuffled polygon on the centroid of a source polygon means that each randomised polygon will correspond to a polygon in the original Set I. The RAO values for each polygon in Set I are ranked along with the RAO values for the corresponding polygons in the randomised partitions, producing a reference distribution for each polygon in turn. One then calculates p-values for each polygon in Set I as described above, dividing the rank by the total number of iterations. As with the global analysis, one compares the p-values for each polygon in Set I to, say $\alpha = 0.05$ or $\alpha = 0.10$, to evaluate whether or not the result is unusual enough to reject the null hypothesis that there is no association between that polygon and those in Set J with which it overlaps.

Results

I applied the RAO analysis to the clustered settlement and environmental variables discussed in the preceding section. I compared each set of polygons created from clustering the settlement data with the polygons generated from clustering the sets of environmental variables chosen through the regression model selection process. For both the global and local RAO tests of significance, I calculated p-values against a reference distribution of N = 499 randomisations of the clustered settlement polygons, giving a total of 500 iterations. Table 19 sets out the results of the global RAO analyses, listing the clustered settlement variables, the clustered environmental variables to which the settlement clusters were compared, and the unweighted and area-weighted RAO statistics and their corresponding p-values.

Table 19: Results of global RAO analyses

Clustered Settlement Variables	Clustered Environmental Variables	Unweighted RAO Value	p-value	Weighted RAO value	p-value
ABCDE (K5)	EnvS1 (K5)	0.137	**0.002**	0.162	0.306
ABCDE (K6)	EnvS1 (K5)	0.126	**0.002**	0.152	0.154
ABCDE (K5)	EnvS1 (K8)	0.167	**0.002**	0.151	0.250
ABCDE (K6)	EnvS1 (K8)	0.143	**0.002**	0.153	**0.050**
BCD (K4)	EnvS2 (K5)	0.130	**0.002**	0.166	0.134
BCD (K6)	EnvS2 (K5)	0.121	**0.002**	0.145	0.242
BCD (K4)	EnvS2 (K7)	0.153	**0.002**	0.135	0.336
BCD (K6)	EnvS2 (K7)	0.149	**0.002**	0.142	0.242
BCD (K4)	EnvS3 (K5)	0.128	**0.002**	0.169	*0.098*
BCD (K6)	EnvS3 (K5)	0.120	**0.002**	0.147	0.180
BCD (K4)	EnvS3 (K11)	0.174	**0.002**	0.135	0.150
BCD (K6)	EnvS3 (K11)	0.181	**0.002**	0.150	*0.088*
CSSNa2 (K5)	EnvS4 (K3)	0.076	**0.002**	0.192	*0.058*
ABCDEDspHC (K3)	EnvS4 (K3)	0.109	**0.002**	0.319	*0.062*
ABCDEDspHC (K10)	EnvS4 (K3)	0.077	**0.002**	0.119	0.978
CSSNa2 (K5)	EnvS4 (K9)	0.149	**0.002**	0.163	*0.100*
ABCDEDspHC (K3)	EnvS4 (K9)	0.178	**0.002**	0.190	**0.032**
ABCDEDspHC (K10)	EnvS4 (K9)	0.160	**0.002**	0.177	*0.054*
CSSNb2 (K5)	EnvS4 (K3)	0.080	**0.002**	0.188	*0.072*
BCDDspHC (K3)	EnvS4 (K3)	0.100	**0.002**	0.326	*0.092*
CSSNb2 (K5)	EnvS4 (K9)	0.142	**0.002**	0.153	0.252
BCDDspHC (K3)	EnvS4 (K9)	0.172	**0.002**	0.173	0.138

P-values in bold are significant at $\alpha = 0.05$; p-values in italic are significant at $\alpha = 0.10$

In all cases, the settlement clusters had the highest average amount of unweighted overlap with their corresponding sets of environmental clusters, giving p-values of 0.002. The unweighted RAO values themselves, however, were all fairly low, averaging 0.135, and none were higher than 0.181. The results for the area-weighted RAO values were more varied. Only two of the 22 analyses showed weighted RAO values significant at $\alpha = 0.05$, and eight more were significant at $\alpha = 0.10$. The weighted RAO values were generally somewhat higher than the unweighted values, averaging 0.173, but in several cases, the weighted RAO values were lower than the unweighted ones. Two cases produced notably higher values, around 0.32: the comparisons of clusters for distance to category A–E and category B–D nucleations combined with dispersion scores and hamlet counts using K = 3 to clusters for environmental variable Set 4 using K = 3. Both cases were significant at $\alpha = 0.10$ but not at $\alpha = 0.05$.

Figure 29–Figure 39 illustrate the results of the local RAO analyses. The polygons of each set of settlement clusters are shaded according to their RAO values, that is, the maximum relative area overlap when compared to the relevant set of clustered environmental variables. For all the maps, I grouped the values into five classes using Jenks's Natural Breaks method (Jenks and Caspall 1971; Jenks 1977) as implemented in ArcGIS (Esri 2012). To aid comparison of the maps, I standardised the legend values across the maps by taking the averages of the cut-off values for the five legend classes. Individual polygons significant at $\alpha = 0.05$ (ie, their p-values are ≤ 0.05) are highlighted in red; those significant at $\alpha = 0.10$ are highlighted in orange. These are the polygons whose RAO values were in the top 5 or 10 per cent when compared to the RAO values for the corresponding polygons in the randomised iterations. Polygons highlighted in dark grey are those whose p-values are higher than 0.95, indicating that the corresponding polygons in more than 95 per cent of the random iterations had an RAO value higher than that found for the original polygon.

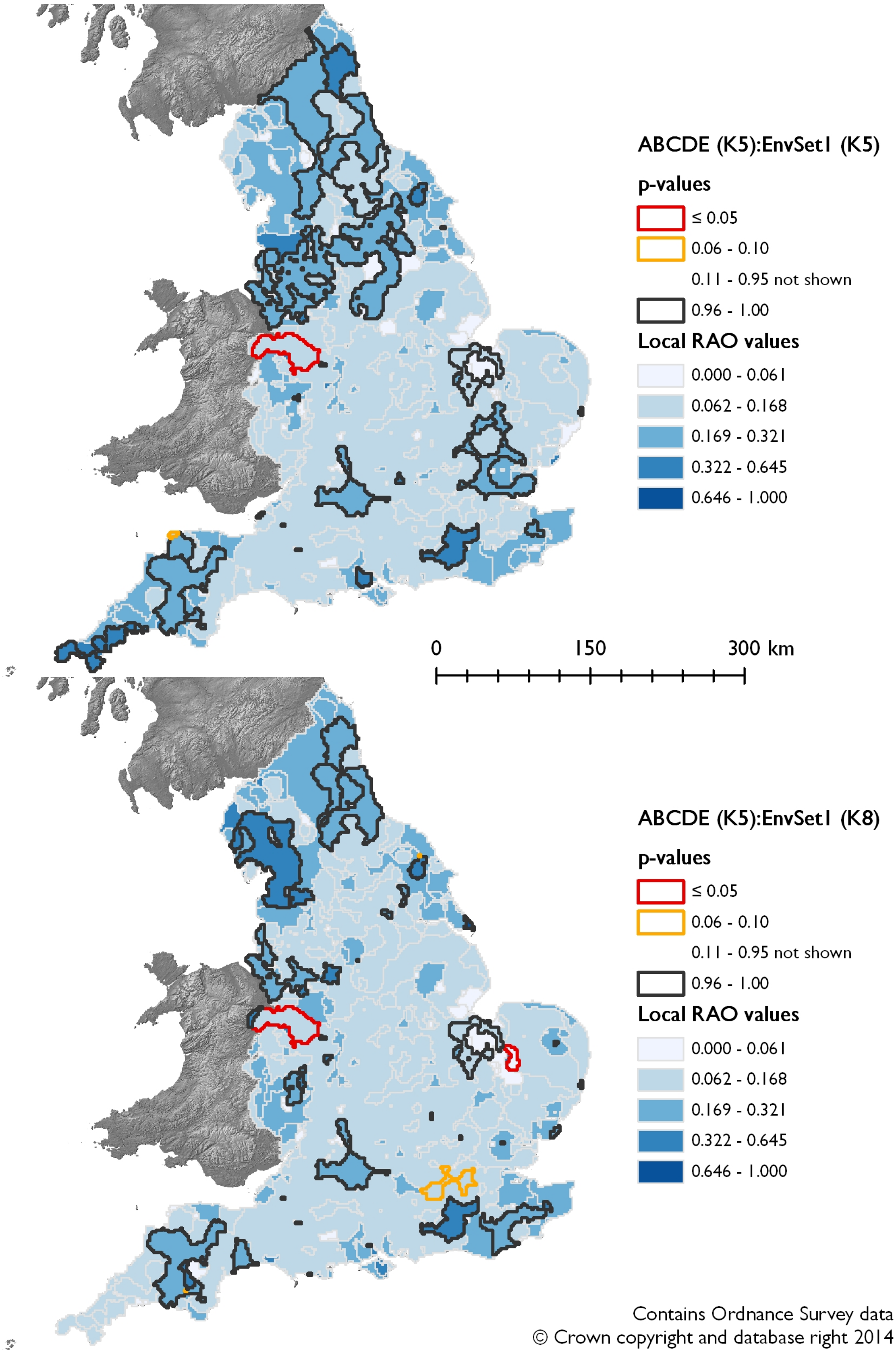


Figure 29: Local RAO results comparing ABCDE (K = 5) to EnvSet 1 (K = 5) (top) and EnvSet 1 (K = 8) (bottom)

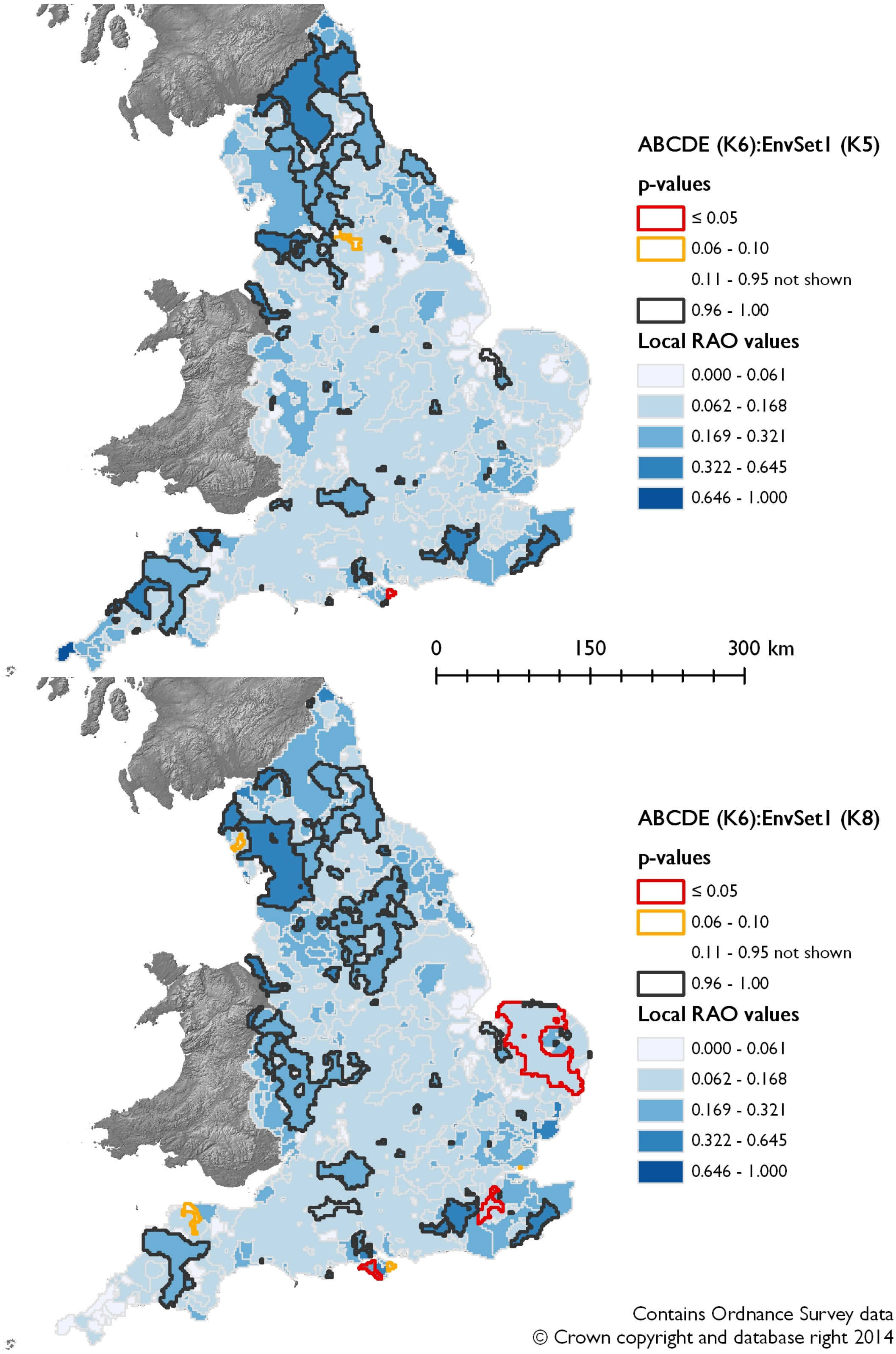


Figure 30: Local RAO results comparing ABCDE (K = 6) to EnvSet 1 (K = 5) (top) and EnvSet 1 (K = 8) (bottom)

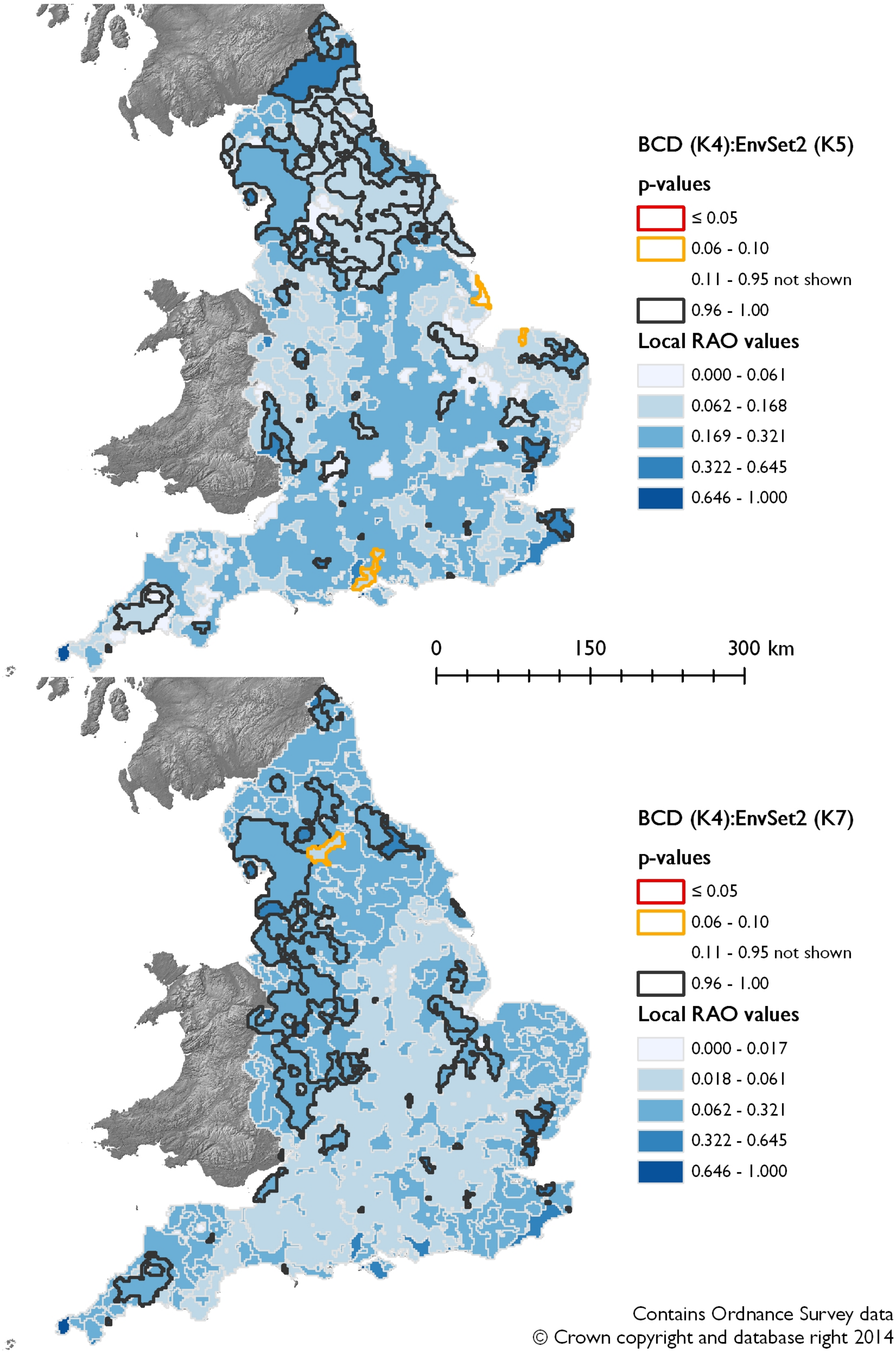


Figure 31: Local RAO results comparing BCD (K4) to EnvSet 2 (K5) (top) and EnvSet 2 (K7) (bottom)

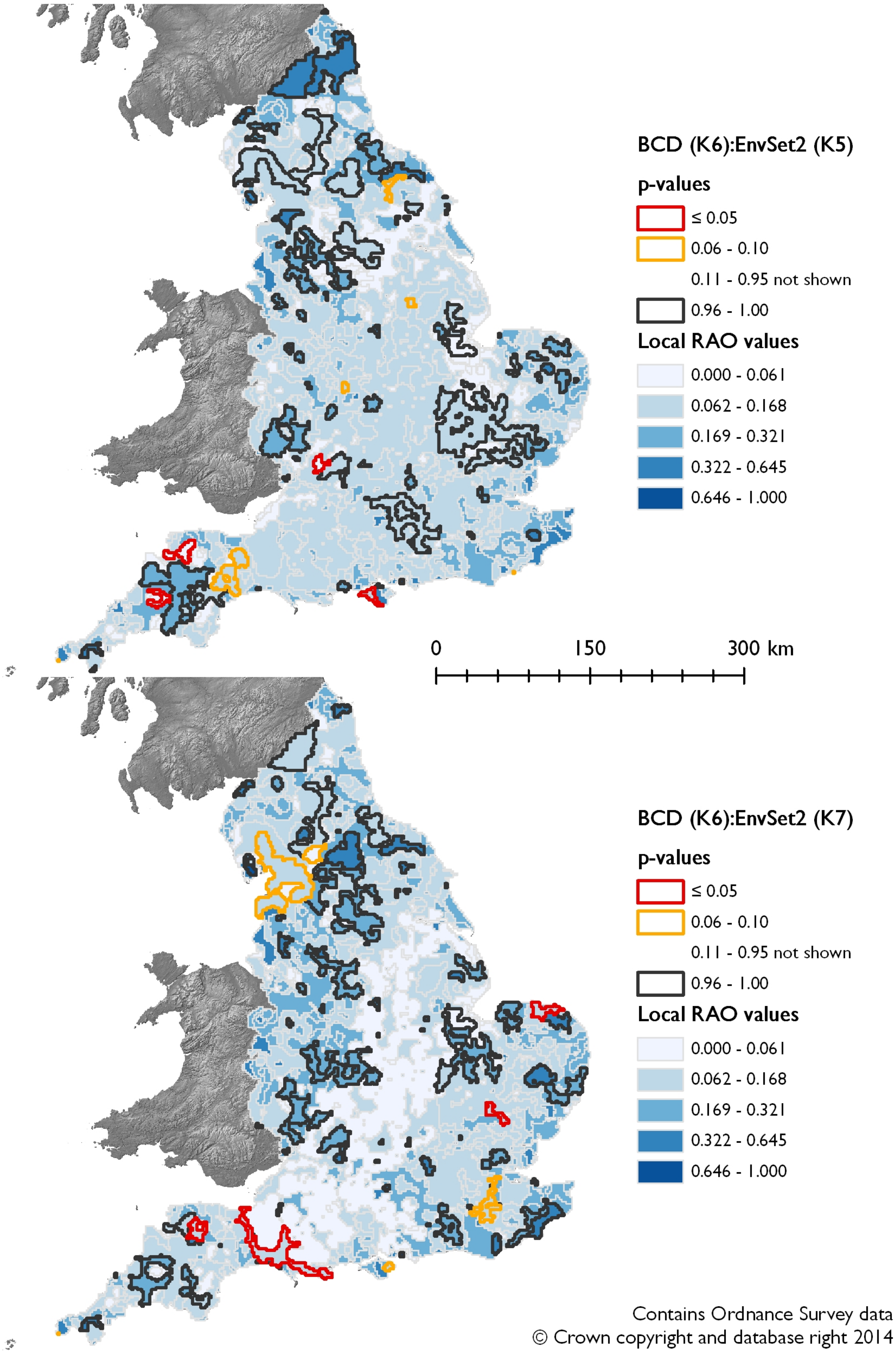


Figure 32: Local RAO results comparing BCD (K = 6) to EnvSet 2 (K = 5) (top) and EnvSet 2 (K = 7) (bottom)

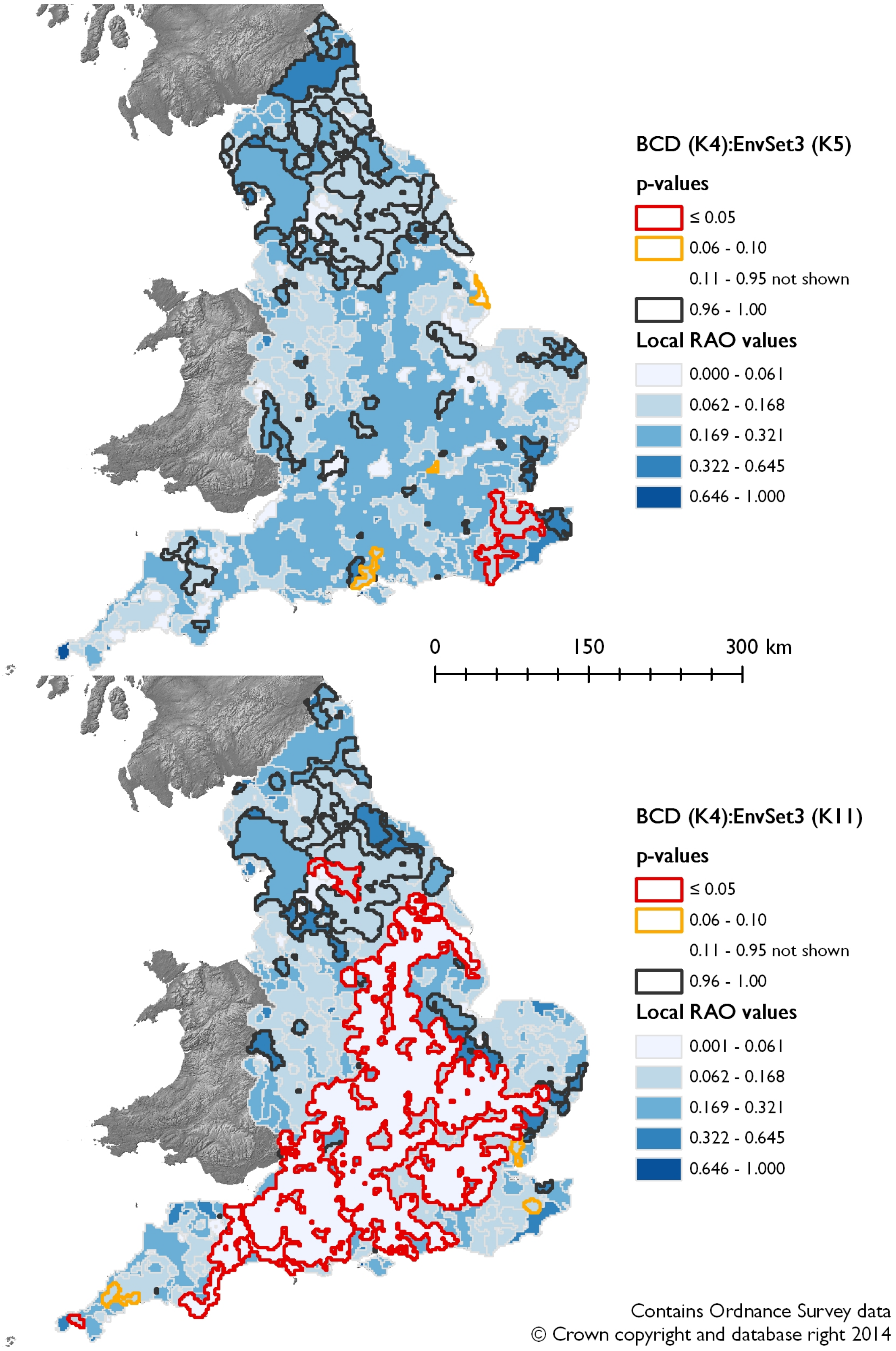


Figure 33: Local RAO results comparing BCD (K = 4) to EnvSet 3 (K = 5) (top) and EnvSet 3 (K = 11) (bottom)

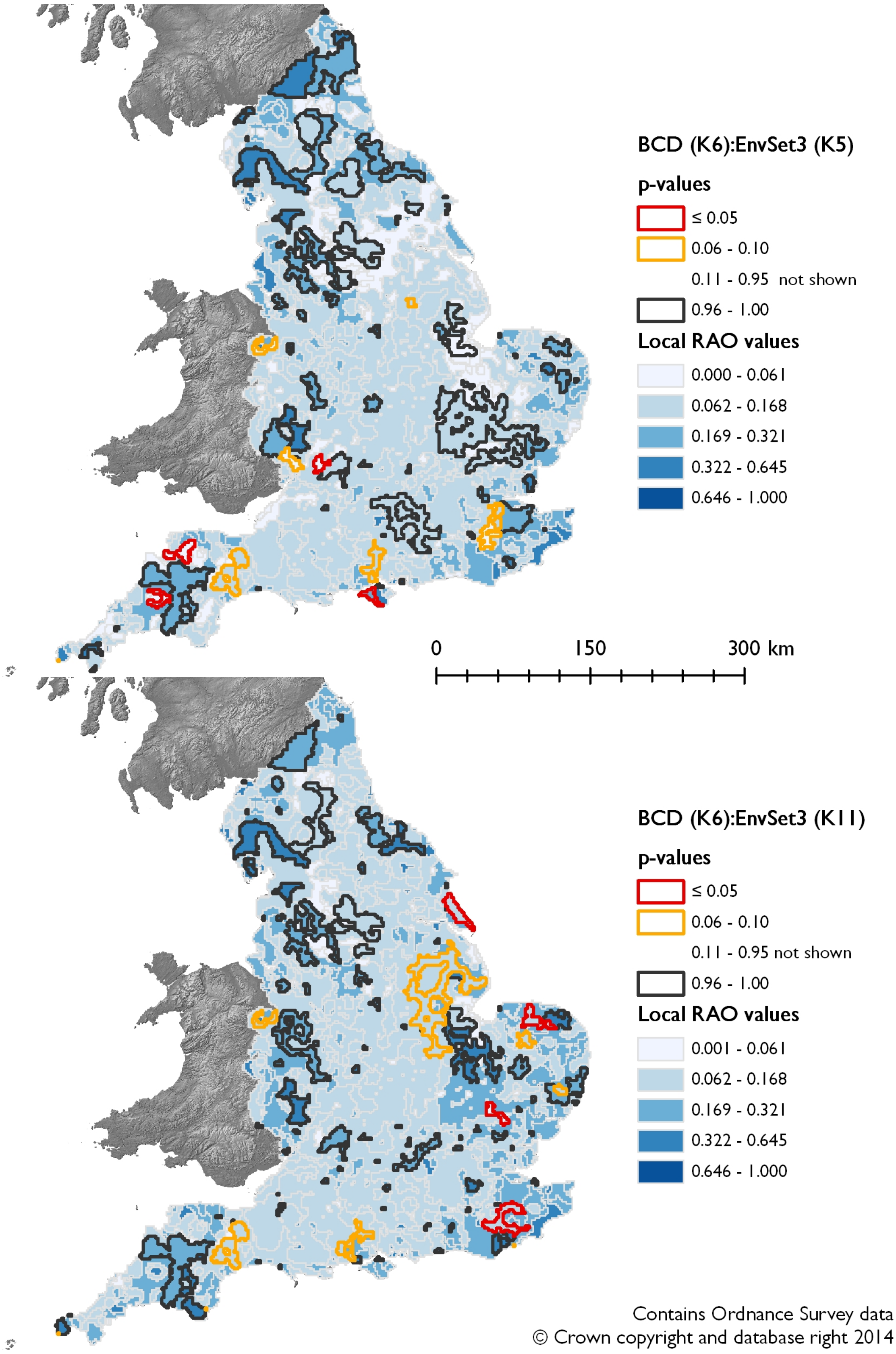


Figure 34: Local RAO results comparing BCD (K = 6) to EnvSet 3 (K = 5) (top) and EnvSet 3 (K = 11) (bottom)

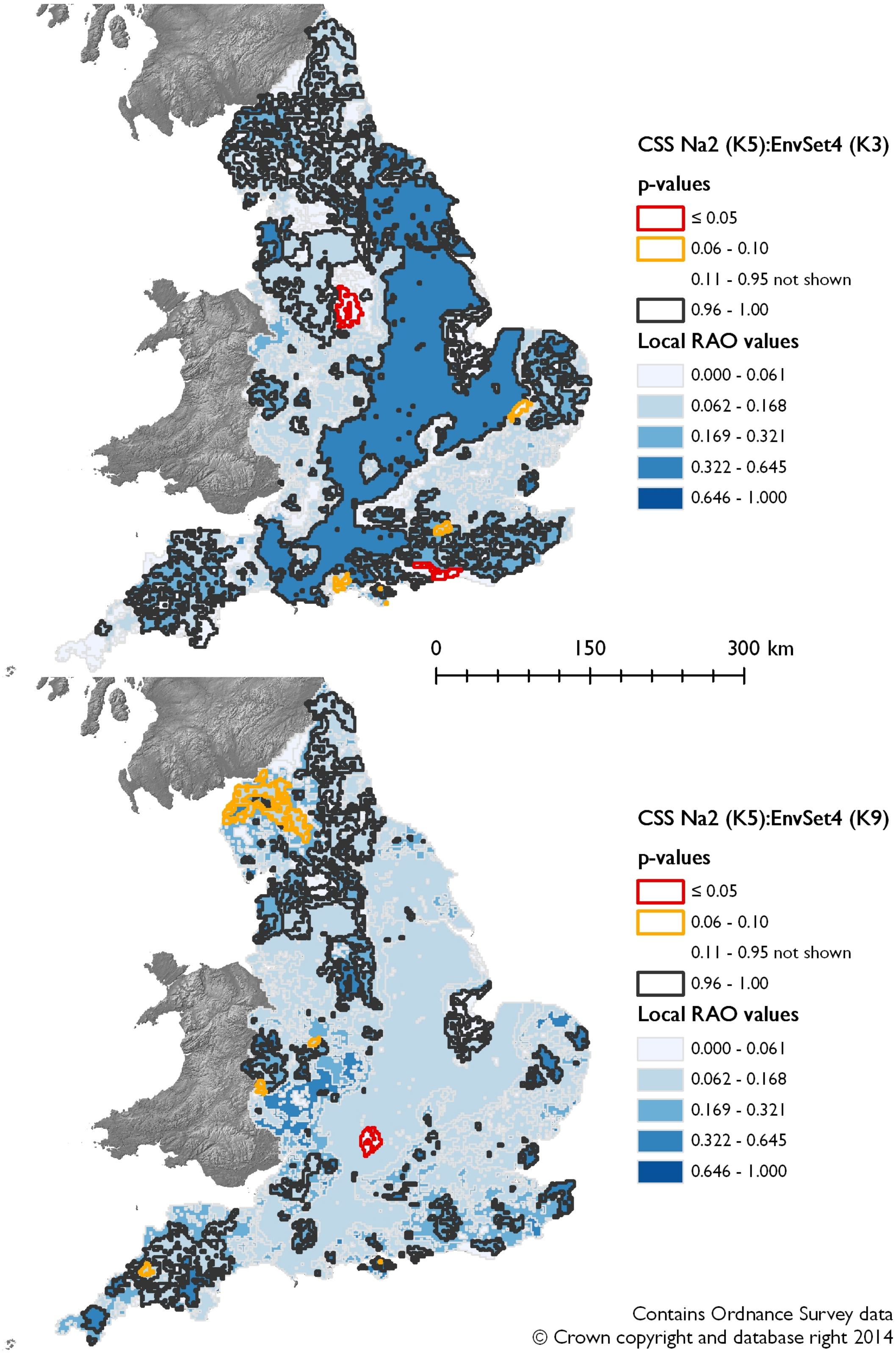


Figure 35: Local RAO results comparing CSS Na2 (K = 5) to EnvSet 4 (K = 3) (top) and EnvSet 4 (K = 9) (bottom)

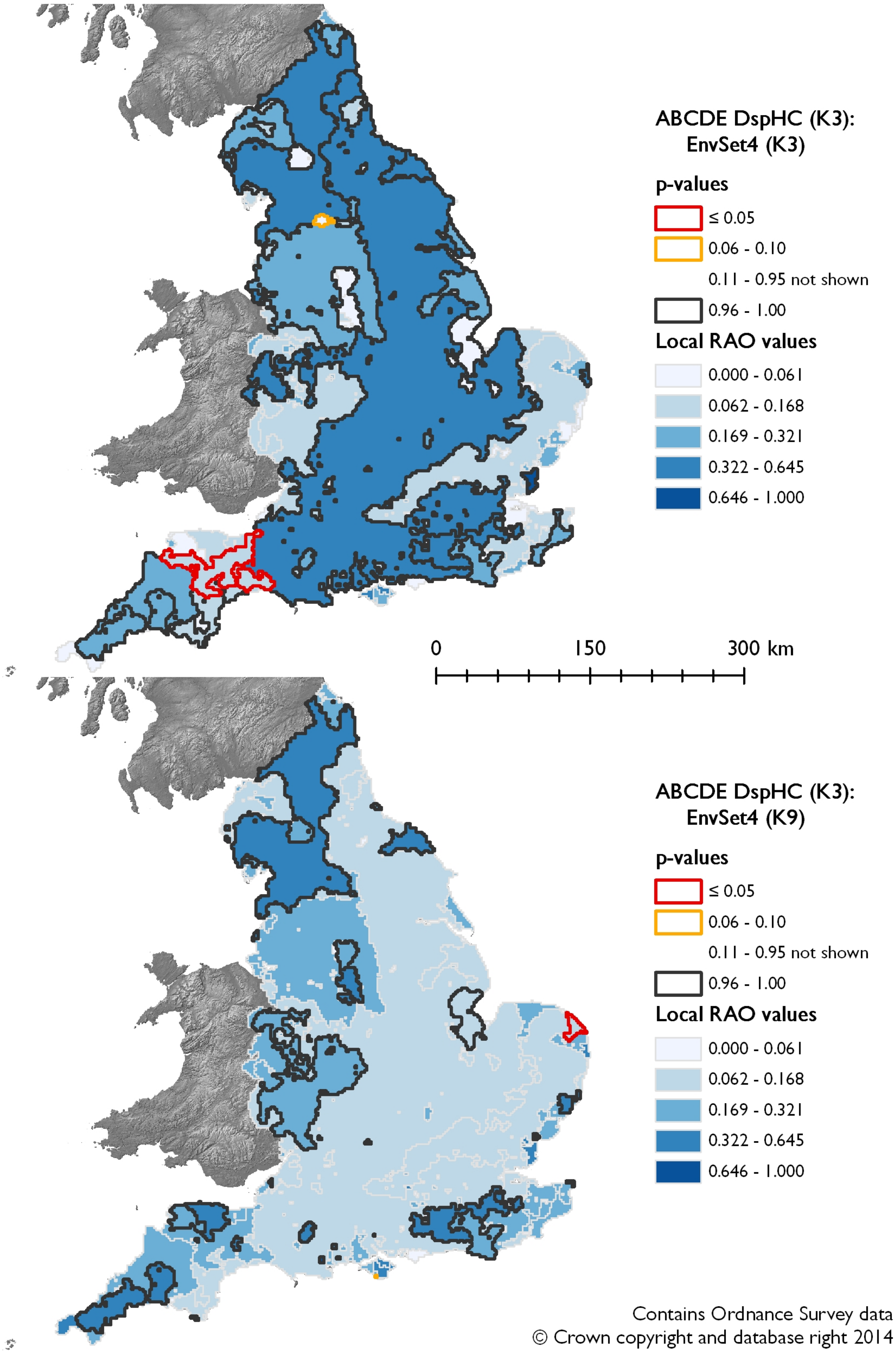


Figure 36: Local RAO results comparing ABCDE DspHC (K = 3) to EnvSet 4 (K = 3) (top) and EnvSet 4 (K = 9) (bottom)

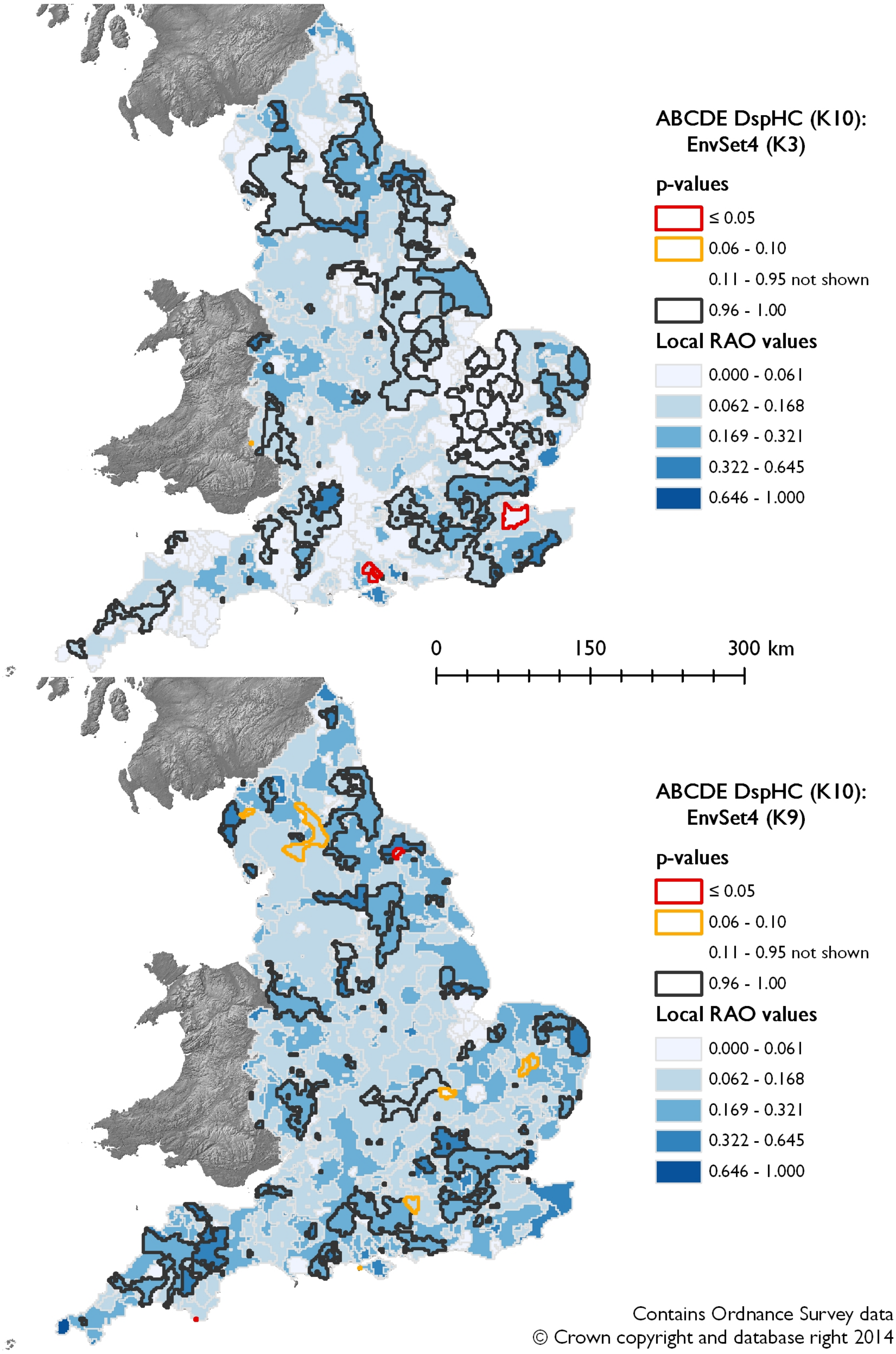


Figure 37: Local RAO results comparing ABCDE DspHC (K = 10) to EnvSet 4 (K = 3) (top) and EnvSet 4 (K = 9) (bottom)

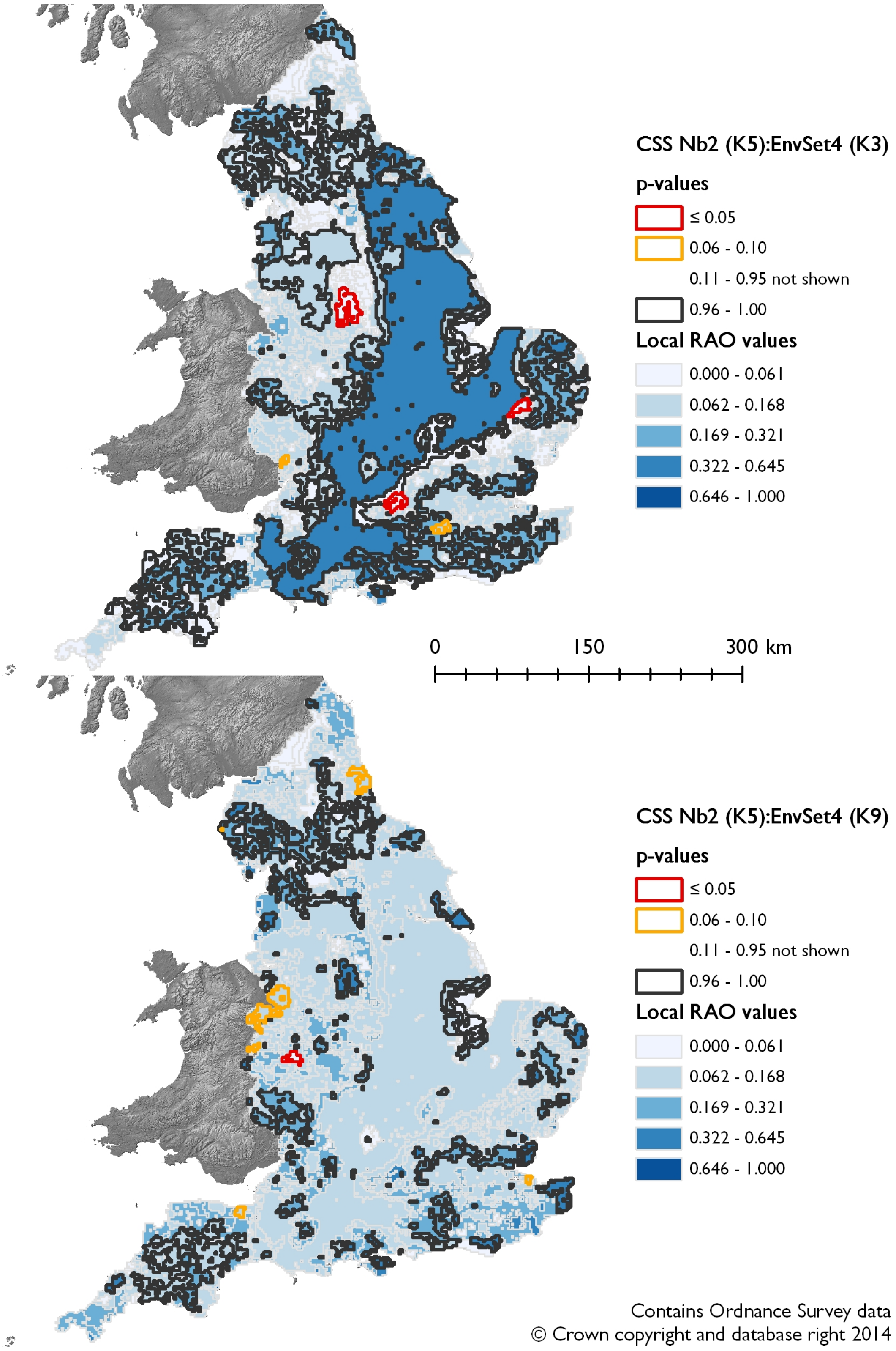


Figure 38: Local RAO results comparing CSS Nb2 (K = 5) to EnvSet 4 (K = 3) (top) and EnvSet 4 (K = 9) (bottom)

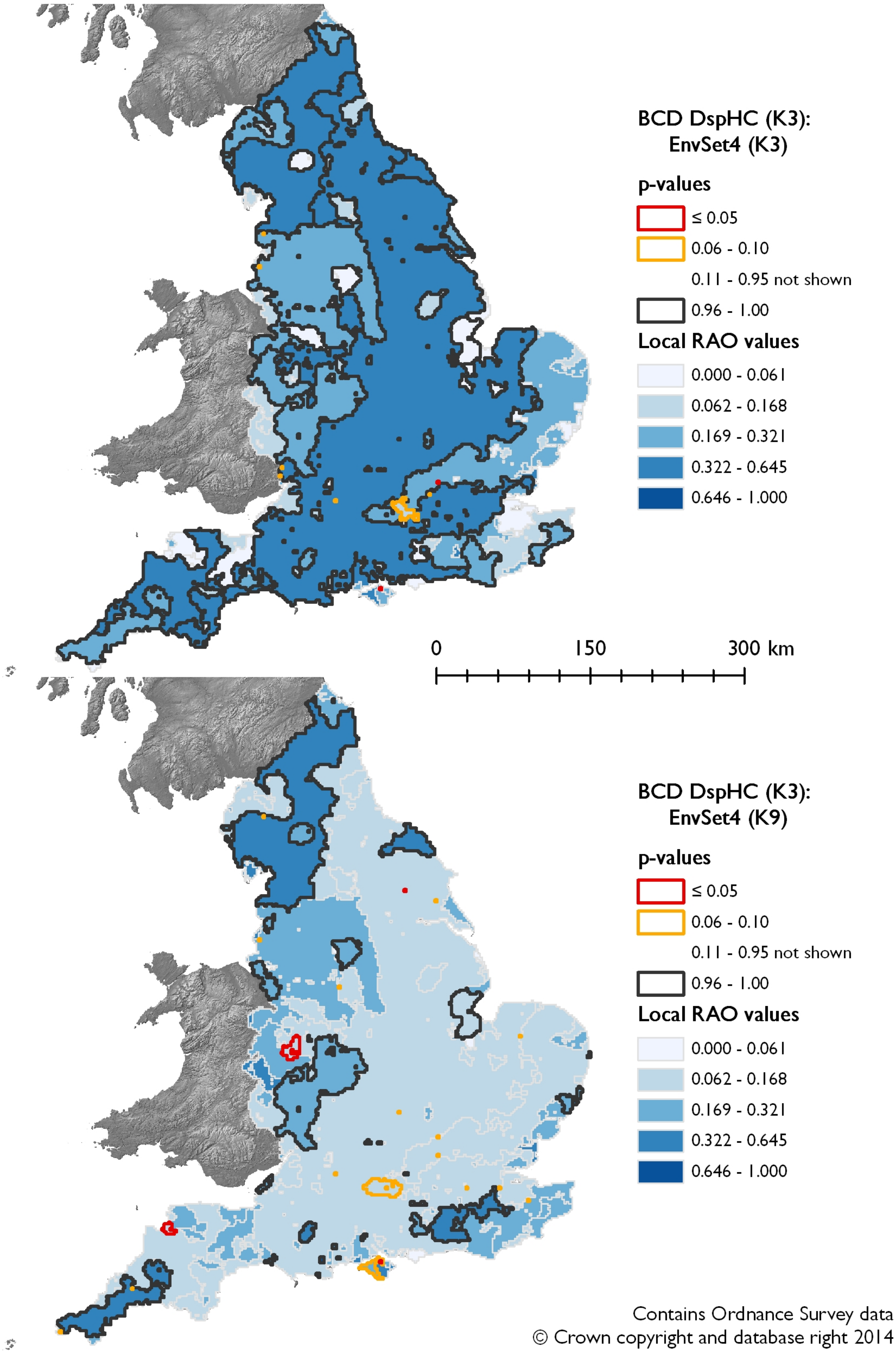


Figure 39: Local RAO results comparing BCD DspHC (K = 3) to EnvSet 4 (K = 3) (top) and EnvSet 4 (K = 9) (bottom)

Discussion

Using both global and local RAO analysis has made it possible to assess in a robust fashion the association or 'degree of match' between clusters derived from the settlement variables on those derived from the environmental variables.

All of the unweighted global RAO values are significant, but these results are almost certainly misleading, most likely due to the inherent limitations of the unweighted statistics relative to the area-weighted statistics (Maruca and Jacquez 2002, 73). As noted above, all the RAO values are fairly low, indicating that even the highest average amount of overlap was quite small. The polygons in the respective partitions are of drastically different sizes, and in all the partitions, there are fairly large numbers of very small polygons, some constituting only a single 2 x 2km cell. The unweighted RAO statistic treats the overlap between a pair of 2 x 2km cells the same as that between a pair of polygons each hundreds of square kilometres in size. For the datasets analysed here, the unweighted statistics give a spuriously optimistic view of the degree of match between the respective sets of polygons.

With two exceptions, the area-weighted global RAO values are also fairly low. When compared to the clusters for environmental variable Set 4 where K = 3 (top of Figure 23), the clusters for distance to category A–E nucleations combined with dispersion scores and hamlet counts also generated using K = 3 (top of Figure 16) had an area-weighted average overlap of 0.319, with a p-value of 0.062. The clusters for distance to category B–D nucleations combined with dispersion scores and hamlet counts using K = 3 (top of Figure 18), similarly compared to those for environmental variable Set 4 where K = 3, had a weighted RAO value of 0.326 with a p-value of 0.092. The weighted RAO values for both these instances are clearly higher than those for any of the other analysed pairs of settlement and environmental clusters, indicating an average amount of maximum area-weighted overlap of about one-third. The p-values are significant at $\alpha = 0.10$, meaning there is about a 1 in 10 chance that rejecting the null hypothesis of no association between the sets of polygons would be an error.

Only two cases produced p-values significant at $\alpha = 0.05$: the comparisons of clusters for distance to category A–E nucleations using K = 6 (top of Figure 11) to environmental variable Set 1 clustered using K = 8 (bottom of Figure 21) and that of, again, clusters for distance to category A–E nucleations combined with dispersion scores and hamlet counts generated using K = 3 (top of Figure 16) to environmental variables set 4 clustered using K = 9 (bottom of Figure 24). The weighted RAO values for both these cases were, however, fairly low: respectively 0.153 and 0.190. The average maximum area-weighted overlap in these cases was highly unusual when compared to randomised data, but the amount of overlap was only about one-sixth to one-fifth – hardly indicative of a close match.

It is worth comparing these results with those Maruca and Jacquez produced as part of their development of the RAO technique using simulated data. The best weighted RAO values for the settlement and environment data are only slightly higher than the weighted RAO value (0.310) for Maruca and Jacquez's test polygons meant to simulate maximal offset between sets, that is, where the boundaries avoid each other as much as possible. The weighted RAO values for three other analyses simulating varying cases of 'good' overlap ranged from 0.444 to 0.833 (Maruca and Jacquez 2002, 78). That having been said, in their the 'real-world' example, Maruca and Jacquez considered that a weighted RAO value of 0.369 (with a corresponding p-value of ≤ 0.002) indicated that partitions overlapped very well (ibid, 80). Other than the two best results, the weighted RAO values for the pairs of settlement and environment polygon sets appear extremely low. In the absence of a more fully-developed range of simulated results to which to compare, it is difficult to say confidently whether the best results from my analyses represent overlap that is good, bad or indifferent. It does seem clear, however, that the overlap between most of the pairs of polygon partitions is poor.

Taking England as a whole, only the simplest partitionings of the settlement nucleation and dispersion data match up reasonably, though not overwhelmingly, well with the simplest partitioning of one set of the environmental factors data. The degree of match seems moderately reliable, though the results are not so unusual as to rule out the possibility that they could simply be the product of chance. Clusters derived from just the nucleation data as well as the more complex partitionings (that is, with higher values for K) of the combined nucleation and dispersion data matched poorly with the corresponding clusters derived from the environmental variables. Even when the results were highly unusual compared to randomised data, the degree of match between the polygon sets was low.

Turning to the local results, in general, few individual settlement cluster polygons were found to be significant at either $\alpha = 0.05$ or 0.10. That is, only a small fraction of the settlement polygons show an exceptional degree of overlap with the environmental factors polygons when compared to the randomised sets. There is almost no consistency in the locations of settlement polygons showing statistically significant degrees of overlap across the sets of results. With very few exceptions, those settlement polygons that do show statistically significant overlap are relatively small and their RAO values are toward the lower end of the spectrum. Only a few of the settlement polygons show a statistically significant high degree of overlap with the environmental factors polygons. Polygons with extremely high p-values (≥ 0.95) were much more common and covered a substantially higher proportion of the country than did those with very low p-values. In other words, randomly generated settlement polygons matched up well with the environmental factors polygons far more often than did the original settlement clusters. Many of the individual polygons with high RAO values also have extremely high p-values. Even where the degree of local overlap is good, such results appear, more often than not, to be the product of chance. This suggests that there is not a meaningful, causative association in those locations between the environmental factors and settlement organisation.

Looking at the maps of the local statistics in turn, a few points are particularly noteworthy. None of the polygons in the clusters for distance to category B–D nucleations using K = 4 were individually significant at $\alpha = 0.05$ when compared against those for environmental Set 2 using K = 5 or 7 (Figure 31), and only a handful were individually significant at $\alpha = 0.10$. In Figure 33, there is one very large polygon extending across much of central and southern England that is significant at $\alpha = 0.05$, but its RAO value is quite low (0.052 to be precise). The p-value for that polygon appears statistically significant not because it overlaps especially well with an environmental polygon, but because the overlap in most of the randomised sets is far worse. The local results for the best-performing cases in the global analysis (category A–E and category B–D nucleations combined with dispersion scores and hamlet counts using K = 3 compared to clusters for environmental variable Set 4 using K = 3) are illuminating. The degree of overlap between the extensive settlement polygons covering large swathes of central and southern England (top of Figure 36 and Figure 39) is good (≥ 0.646). The p-values for these polygons, however, are also very high (≥ 0.95), indicating that overlap with the environmental polygons is better in the overwhelming majority of corresponding polygons in the randomised sets. The degree of overlap is quite good, but it appears highly likely that rejecting the null hypothesis of no association would be an error.

It must be noted, however, that the local statistics presented here are unweighted, and so may suffer from similar issues as the unweighted global statistics. Again, the unweighted RAO statistic treats the overlap between a pair of 2 x 2km cells the same as that between a pair of much larger polygons. A good match between a large settlement polygon in the original data and an environmental polygon should, arguably, carry more interpretative weight than a match between a small randomised settlement polygon and an overlapping environmental polygon. The settlement polygons in the randomised sets means can be of varying sizes when compared to the corresponding polygons in the original source set. That is, a polygon in the original set with the unique ID 101 might have an area of 45.67km^2. In a randomised set, the polygon with unique ID 101 might only have an area of 4.56km^2, and in another randomised set, polygon 101 might have an area of 456.78km^2. A good, but less than perfect, degree of overlap between the original settlement polygon and its best match in the environmental set ought to be given greater statistical weight than 100% overlap between the much smaller polygon 101 in the first randomised set and its best match environmental polygon. And, of course, a good amount of overlap using the much larger polygon in the second hypothetical random set ought to count for even more than the overlap found using the original source polygon. Without area-weighting, the statistics computed here do not reflect these possibilities. It may be the case that the unweighted local statistics give an unduly pessimistic view of the degree of match between the individual polygons in the respective sets. Further work to implement an area-weighted version of the local statistics is clearly required.

On a global level, then, the broadest patterns discerned in the settlement nucleation and dispersion data seem to follow, roughly, the broadest patterns derived from one set of environmental variables. This general correspondence has long been recognised, and, indeed, it would be somewhat surprising if the results of the global RAO analysis did not reflect it. Beyond the general trend, however, the association between settlement and environmental patterns breaks down. More complex patterns extracted from the settlement data do not appear to match up well with those apparent in the environmental variables. Purely visual comparison of the maps of settlement and environmental clusters suggests that in some cases, the boundaries – that is the lines dividing one polygon from another – in the settlement data match up with some of the boundaries in the environmental data, but the areas of the clusters, on the whole, do not.

From a methodological standpoint, it is clear that Maruca and Jacquez's RAO technique can be a useful tool for evaluating associations between phenomena represented as polygons. The local version of RAO can show where overlaps are good and where they are not, providing greater insight into patterns of association than do the global statistics. Further work is, however, required to develop an area-weighted version of the local statistics. Having used this method, the association between the clustered environmental factors and the clustered historic settlement data appears weak. These results do not suggest that patterns of soils, precipitation, temperature and elevation strongly influenced patterns of historic settlement organisation across England.

CONCLUSIONS

The first – admittedly rather pedestrian – point to be made is that the work presented here demonstrates that the project's research aims and objectives have successfully been met. It is possible to collate GIS data for historic settlement nucleation and dispersion with a range of data on environmental variables. The regression model specification, selection and validation processes, followed by further analysis using spatial regression methods, identified environmental variables that appear to have had the most significant influence on regional variation in historic settlement organisation. The use of both non-spatial and spatial regression and the Relative Area Overlap Analysis has enabled investigation of how relationships between key environmental variables and historic settlement organisation varied across England. Using the unsupervised classification clustering technique, it has been possible to develop new, national-scale characterisations of historic settlement organisation and of key environmental variables. These new classifications of historic settlement organisation often broadly align with Roberts and Wrathmell's delineations of provinces, sub-provinces and local regions, but the cluster outlines and Roberts and Wrathmell's boundaries diverge more often than they agree.

The kinds of rigorous spatial statistical analysis presented here have only been possible because of the conversion of Roberts and Wrathmell's maps to GIS data, along with the large, and ever-growing, range of GIS-ready environmental data. This work has also demonstrated that GIS and related quantitative spatial analytical tools can be used to investigate a key topic in English landscape history and archaeology in sophisticated ways, moving well beyond using GIS for simple data management and map production. This research has built on English Heritage's investment in both the original *Atlas* and in the GIS dataset produced from it and demonstrates the potential embodied in the Atlas of Rural Settlement in England GIS dataset.

What, then, has this study revealed about the relationships between environmental factors and rural settlement organisation? The analyses have centred on testing Williamson's explanatory model for the development of settlement nucleation, a model which focuses on the influence of environmental factors such as soils, precipitation, topography and the availability of land suitable for creating meadow. The OLS regression model specification, selection and validation processes identified a number of different sets of environmental variables that fit the data well, but there is a degree of uncertainty as to which model specifications perform best. It is noteworthy that the more complex models performed unequivocally better than simpler models. Some, but by no means all, of the coefficients in the OLS models agree with Williamson's explanatory model.

Incorporating spatial effects using spatial regression methods dramatically improves the performance of the best models. The results suggest that both endogenous processes, eg, the diffusion of nucleation as an approach to settlement organisation, and spatially structured missing variables contributed substantively to variation in rural settlement organisation, in addition to the environmental factors specified in the models. Having controlled for spatial autocorrelation in either the response variables or the models' error terms, many of the soils variables – including several that Williamson argues were influential in the development of nucleated settlement – become insignificant. Much of the variation in the settlement variables is accounted for by the spatially autoregressive parameters, not by the environmental variables.

Overall, the regression analyses indicate that far more of the variation in the measures of settlement organisation is not explained by the environmental variables than is explained by them. The results of the RAO analysis echo this conclusion. At best, the broad spatial patterns extracted from the settlement data roughly follow the broad spatial patterns derived from the environmental variables, but beyond the general trend, the association between settlement and environmental patterns appears weak.

This is not to say that Williamson is wrong to highlight the importance of environmental influences on the development of rural settlement and land-use. Williamson's efforts in trying to pull the pendulum of interpretation back from its pronounced swing in the 'cultural' direction are undoubtedly worthwhile. The goal here has been to illuminate the importance of environmental factors relative to other, unspecified factors, which are assumed – largely for convenience's sake – to be 'cultural' in nature. Environmental data amenable to quantitative analysis together with the settlement data are readily available; data on likely 'cultural' factors generally are not, or at least not on a national scale.

There are, of course, unresolved issues in the models, which make any conclusions based on them open to question. The diagnostics for both the non-spatial and spatial regression models indicate problems of mis-specification, problems which are, on one level, to be expected. It might be argued that, given the imperfect nature of the regression results, the outcomes of the clustering analyses are flawed because they are based on unreliable sets of environmental variables. Continuing the argument, the RAO test results may also be flawed because they use the clustered environmental and settlement data as inputs. Such critique may well be valid. My conclusions should not be taken as definitive, but rather a further, contestable contribution to scholarship on the nature and causes of variation in rural settlement organisation in England. Further work, to improve and refine the methods and results set out here, is very much desirable.

I will make two further points of reflexive critique. First, I recognise that I have used very simple – even crude – measures of rural settlement organisation, which do not capture the true richness of variability in rural settlement form and function. The landscape of nucleated settlement in Durham is not identical to the landscape of nucleated settlement in Dorset. Likewise, highly dispersed settlement in Cumbria is not the same as highly dispersed settlement in Cornwall. This highlights the problem of equifinality – different processes or sets of variables in different places can produce what is measured as the same outcome, whether distance to nearest nucleation or combined settlement score. The only way to address this issue, however, would be to collect better, more detailed data on historic rural settlement.

Second, it is clear that this has been an analysis of data depicting nineteenth-century settlement organisation, based on a snapshot of a specific moment in the long (and on-going) process of rural settlement evolution. I do not claim to have analysed directly the development over time of medieval and later rural settlement. This point draws attention to one of the basic challenges of any cross-sectional analysis: inferring and interpreting complex processes that operated over both time and space from a single 'snapshot' of data is fraught with difficulty. There is no easy solution to this problem, but it is one with which all archaeologists and historians are familiar. Roberts and Wrathmell were confident that the patterns in rural settlement they detected in the nineteenth-century mapping could, in broad terms, be projected further back into the past (Roberts and Wrathmell 2000, 27-37). Their arguments for the antiquity of such patterns suggest that the results of the analyses presented here can also be projected back in time, but there can be no genuine certainty as to their validity. Only more work, based on data depicting rural settlement in earlier time periods than Roberts and Wrathmell's, could address this issue directly.

Turning to the future, as noted in the individual analysis sections, there are numerous possibilities for further work, building upon the data and methods presented here. This study has focused exclusively on investigating relationships between measures of settlement organisation and environmental influences, but field systems and land-use are, arguably, equally important aspects of landscape character. There are, to my knowledge, no freely-available data that are national in scope on historic field systems and land-use covering the same period as Roberts and Wrathmell's settlement data. Historic Landscape

Characterisation (HLC) data on fieldscapes and land-use (Turner 2006) might be used to add greater depth and complexity to the measures of settlement organisation. HLC data are not without their own issues (Williamson 2007), and coverage of England is not yet complete. The potential of comprehensive HLC data for England to enhance a study like this one appears tantalising, but remains, as yet, unexplored.

Other data sources for soils could prove more amenable to the kinds of analysis attempted here. Possible alternatives to the Soilscapes data include the full National Soil Map of England and Wales (National Soils Research Institute (NSRI) 2014a) and the National Soils Inventory (National Soils Research Institute (NSRI) 2014b), as well as the British Geological Survey (BGS) Soil Parent Material database (Lawley 2009). These datasets are not freely available to all parties, so any possible future work using these data would carry serious, potentially prohibitive, cost implications. Future work could also explore the influence of the underlying bedrock geology and on superficial deposits (also known as 'drift geology'), for example using BGS's DiGMapGB-625 dataset (British Geological Survey 2008; 2003). These data are relative low-resolution (a nominal scale of 1:625 000), but they are freely available. Whether bedrock and superficial geology might augment or supplant soils data for analyses such as those presented here must remain an open question.

It is well known that 'the greater the level of spatial detail the higher the level of noise there is likely to be in the data and the greater the need to draw on methods that distinguish between "noise" and "signal" in pattern analysis and in the analysis of relationships' (Haining 2003, 42). The fine 'grain' of the data used here (2 × 2km grid squares) may contain a great deal of 'noise' which could be obscuring meaningful relationships operating over larger areas. It would be technically straightforward, if time-consuming, to resample all the data used in this study to coarser resolutions, eg, 5 × 5, 10 × 10 or 20 × 20km squares, and apply the same analytical approaches described above. Doing so would allow investigation of whether environmental factors influenced settlement organisation at lower levels of granularity and, even more intriguingly, whether different factors may have operated at spatial resolutions.

Further work could explore potential remedies for the kinds of regression model issues highlighted by the various diagnostic tests, such as transformations of the response and explanatory variables, and the use of spatial trend and interaction variables, as well as additional spatial regression methods. Given the problems involved in regression analysis (both aspatial and spatial), it would also be worth exploring alternative methods to explore associations between environmental factors and settlement organisation, methods less restricted by rigid assumptions about the data.

Geographically Weighted Regression (GWR) (Fotheringham *et al* 2002) could be used to investigate the spatial variation in the strength of the relationships between those environmental variables identified as meaningful using OLS and spatial regression and historic settlement nucleation and dispersion. GWR models the extent to which statistical relationships may differ from place to place, rather than assuming that relationships between variables were constant across the whole area studied, as do OLS and spatial regression. GWR is specifically designed to explore relationships which are non-stationary. It is clear that the models selected using OLS and spatial regression are mis-specified, that is, there are explanatory variables that have not been included, so it is unclear whether GWR would produce reliable results. Given how GWR works, collinear data can cause both computational and interpretative problems (Wheeler 2007; Wheeler and Calder 2007; Wheeler and Tiefelsdorf 2005), and there are not, as yet, any straightforward ways to deal with the issues (Brunsdon *et al* 2012). Many of the variables used in this study – those for soils especially – are highly collinear, so the application of GWR could be problematic.

Another issue with the regression analyses performed here, as well as with GWR, is that all these techniques assume that the relationships between the explanatory and response variables are linear. A recently-developed method, the Local Entropy Map (Guo 2010), could be useful, in that it does not assume that relationships between variables are all linear, and it explicitly considers the spatial location of observations when analysing multivariate relationships.

Turning to the clustering analysis, generalising or smoothing the clusters produced by the ISODATA unsupervised classification would reduce the spatial fragmentation – the 'speckled' appearance – of the spatial outputs of the process (Lillesand *et al* 2008, 580-1). There are other approaches to clustering both the settlement data and the various environmental datasets, in particular methods that explicitly incorporate the spatial locations of the observations being clustered, which would very likely produce meaningfully different results. Spatial clustering or region-building algorithms such as SKATER (Assunçao *et al* 2006), REDCAP (Guo 2008; Guo and Wang 2011) and Max-p (Duque *et al* 2012), as well as multivariate clustering and geovisualisation methods built using Self-Organising Maps (SOM) (Kohonen 2001; Guo *et al* 2005; Guo *et al* 2006; Gonçalves *et al* 2008) would appear to hold the most promise. It would also be possible to incorporate uncertainty into the classification/clustering process using fuzzy logic (Burrough and McDonnell 1998, 265-91), recognising that boundaries in data are not always crisp and membership of groups is not always clear-cut.

Employing alternative approaches to clustering or classifying both the environmental and settlement data, of course, has implications for using the RAO method. Different sets of clusters will obviously produce different results when they are overlapped. As noted above, using an area-weighted version of the local RAO statistics could account for the widely varying sizes of the clustered polygons.

Formal analysis of the overlap of boundaries – zones of rapid change in the intensity of variables over geographic space – could also help illuminate associations between environmental factors and settlement organisation. A variety of boundary detection techniques exist (Oden *et al* 1993; Csillag *et al* 2001; Fortin and Dale 2005, 184-202), and the boundary overlap tests developed by Jacquez (1995) have been used in ecological and epidemiological studies to evaluate relationships between distributions of species or disease incidence and an array of environmental variables (Fortin *et al* 1996; Jacquez *et al* 2008; Jacquez 2010). Comparing the boundaries detected in the settlement and environmental data could complement the analysis of the overlap of clustered areas presented here.

Finally, it is worth noting the point made by the statistician George Box that 'all models are wrong; the practical question is how wrong do they have to be to not be useful' (Box and Draper 1987, 74). Are the models described here so wrong as to not be useful? I do not believe they are. Starting from Williamson's explanatory, prose-and-map-based model, I have shown how it is possible to bring together a range of spatial data and quantitative techniques to explore relationships between environmental factors and rural settlement. The models developed here enable the systematic evaluation of a range of hypotheses about those relationships, and to assess the extent to which Williamson's model holds when applied to the whole of England. There is great scope for improvement on the data, methods and models used here, but this study has, I think, demonstrated the potential of GIS-based statistical and spatial analytical approaches in the study of historic settlement studies in England. My hope is that future research on variation in historic settlement organisation in England, building on the work described here, may be able to 'disentangle the relative effects of ecological *vs* cultural factors' (Manning *et al* 2013a, 1046) in a comprehensive, rigorous fashion.

REFERENCES

Akaike, H 1973 'Information Theory and an Extension of the Maximum Likelihood Principle', in Petrov, B N and Caski, F (eds) *Proceeding of the Second International Symposium on Information Theory*. Budapest: Akademiai Kiado, 267-281

Aldrich, J 2006 'When are inferences too fragile to be believed?'. *Journal of Economic Methodology* **13**(2), 161-177

Anselin, L 1988 *Spatial Econometrics: Methods and Models* (Studies in Operational Regional Science **4**). Boston and Dordrecht: Kluwer Academic Publishers

Anselin, L 2002 'Under the hood issues in the specification and interpretation of spatial regression models'. *Agricultural Economics* **27**(3), 247-267

Anselin, L 2005 *Exploring Spatial Data with GeoDa™: A Workbook*. Urbana, IL: University of Illinois, Urbana-Champaign

Anselin, L 2009 'Spatial Regression', in Fotheringham, A S and Rogerson, P A (eds) *The SAGE Handbook of Spatial Analysis*. London: SAGE Publications, 254–75

Anselin, L and Bera, A K 1998 'Spatial Dependence in Linear Regression Models with and Introduction to Spatial Econometrics', in Ullah, A and Giles, D (eds) *Handbook of Applied Economic Statistics* (Statistics: Textbooks and Monographs **155**). New York: Marcel Dekker, 237-289

Anselin, L, Bera, A K, Florax, R and Yoon, M J 1996 'Simple diagnostic tests for spatial dependence'. *Regional Science and Urban Economics* **26**(1), 77-104

Anselin, L, Syabri, I and Kho, Y 2006 'GeoDa: An Introduction to Spatial Data Analysis'. *Geographical Analysis* **38**(1), 5-22

Assunçao, R M, Neves, M C, Câmara, G and Da Costa Freitas, C 2006 'Efficient regionalization techniques for socioeconomic geographical units using minimum spanning trees'. *International Journal of Geographical Information Science* **20**(7), 797-811

Bailey, T C and Gatrell, A C 1995 *Interactive Spatial Data Analysis*. Harlow: Longman Scientific & Technical

Ball, G H and Hall, D J 1965 *ISODATA, a novel method of data analysis and pattern classification*. DTIC Document

Banham, D 2010 '"In the sweat of thy brow shalt thou eat bread": cereals and cereal production in the Anglo-Saxon landscape', in Higham, N J and Ryan, M J (eds) *The Landscape Archaeology of Anglo-Saxon England*. Woodbridge: Boydell & Brewer, 175–192

Bartlein, P, Harrison, S, Brewer, S, Connor, S, Davis, B, Gajewski, K, Guiot, J, Harrison-Prentice, T, Henderson, A, Peyron, O, Prentice, I, Scholze, M, Seppä, H, Shuman, B, Sugita, S, Thompson, R, Viau, A, Williams, J and Wu, H 2011 'Pollen-based continental climate reconstructions at 6 and 21 ka: a global synthesis'. *Climate Dynamics* **37**(3), 775–802

Beheim, B A and Bell, A V 2011 'Inheritance, ecology and the evolution of the canoes of east Oceania'. *Proceedings of the Royal Society B: Biological Sciences* **278**(1721), 3089-3095

Belsley, D A 1991 *Conditioning Diagnostics: Collinearity and Weak Data in Regression* (Wiley Series in Probability and Statistics. Hoboken, NJ: John Wiley & Sons, Inc.

Bevan, A 2012 'Spatial methods for analysing large-scale artefact inventories'. *Antiquity* **86**, 492–506

Bevan, A and Conolly, J 2009 'Modelling spatial heterogeneity and nonstationarity in artifact-rich landscapes'. *Journal of Archaeological Science* **36**(4), 956–964

Bini, L M, Diniz-Filho, J A F, Rangel, T F L V B, Akre, T S B, Albaladejo, R G, Albuquerque, F S, Aparicio, A, Araújo, M B, Baselga, A, Beck, J, Isabel Bellocq, M, Böhning-Gaese, K, Borges, P A V, Castro-Parga, I, Khen Chey, V, Chown, S L, De Marco, J P, Dobkin, D S, Ferrer-Castán, D, Field, R, Filloy, J, Fleishman, E, Gómez, J F, Hortal, J, Iverson, J B, Kerr, J T, Daniel Kissling, W, Kitching, I J, León-Cortés, J L, Lobo, J M, Montoya, D, Morales-Castilla, I, Moreno, J C, Oberdorff, T, Olalla-Tárraga, M Á, Pausas, J G, Qian, H, Rahbek, C, Rodríguez, M Á, Rueda, M, Ruggiero, A, Sackmann, P, Sanders, N J, Carina Terribile, L, Vetaas, O R and Hawkins, B A 2009 'Coefficient shifts in geographical ecology: an empirical evaluation of spatial and non-spatial regression'. *Ecography* **32**(2), 193-204

Boots, B and Okabe, A 2007 'Local statistical spatial analysis: Inventory and prospect'. *International Journal of Geographical Information Science* **21**(4), 355–375

Box, G E P and Draper, N R 1987 *Empirical Model Building and Response Surfaces*. New York: John Wiley & Sons

Brewer, C A and Pickle, L 2002 'Evaluation of Methods for Classifying Epidemiological Data on Choropleth Maps in Series'. *Annals of the Association of American Geographers* **92**(4), 662–681

Brewer, S, Guiot, J and Torre, F 2007 'Mid-Holocene climate change in Europe: a data-model comparison'. *Climate of the Past* **3**(3), 499–512

British Geological Survey 2003 *Digital Geological Map of Great Britain 1:625 000 scale (DiGMapGB-625), superficial deposits data. Version 1.10* [Computer file]. Keyworth, Nottingham: British Geological Survey. Last updated 30 April 2003.

British Geological Survey 2008 *Digital Geological Map of Great Britain 1:625 000 scale (DiGMapGB-625), bedrock data. Version 5.17* [Computer file]. Keyworth, Nottingham: British Geological Survey. Last updated 11 February 2008.

Brunsdon, C, Charlton, M and Harris, P 2012 'Living with Collinearity in Local Regression Models', in Vieira, C, Bogorny, V and Aquino, A R (eds) *Accuracy 2012: Proceedings of the Tenth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences. 10-13th July 2012, Florianópolis, SC, Brazil*. Florianópolis, SC, Brazil: International Spatial Accuracy Research Association and Federal University of Santa Catarina, 67-72

Büntgen, U, Tegel, W, Nicolussi, K, McCormick, M, Frank, D, Trouet, V, Kaplan, J O, Herzig, F, Heussner, K-U, Wanner, H, Luterbacher, J and Esper, J 2011 '2500 Years of European Climate Variability and Human Susceptibility'. *Science* **331**(6017), 578–582

Burnham, K P and Anderson, D R 2002 *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd edn. New York: Springer

Burnham, K P and Anderson, D R 2004 'Multimodel Inference: Understanding AIC and BIC in Model Selection'. *Sociological Methods & Research* **33**(2), 261-304

Burrough, P A and McDonnell, R A 1998 *Principles of Geographical Information Systems*, 2nd edn. Oxford: Oxford University Press

Burt, J E, Barber, G M and Rigby, D L 2009 *Elementary Statistics for Geographers* 3rd edn. New York: Guilford Press

Calinski, T and Harabasz, J 1974 'A dendrite method for cluster analysis'. *Communications in Statistics* **3**(1), 1-27

Chang, K-T 2012 *Introduction to Geographic Information Systems*, 6th edn. Boston: McGraw-Hill

Chatfield, C 1995 'Model Uncertainty, Data Mining and Statistical Inference'. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **158**(3), 419-466

Cliff, A D and Ord, J K 1973 *Spatial Autocorrelation* (Monographs in Spatial and Environmental Systems Analysis **5**). London: Pion

Cliff, A D and Ord, J K 1981 *Spatial Processes: Models and Applications*. London: Pion

Conolly, J and Lake, M 2006 *Geographical Information Systems in Archaeology*. Cambridge: Cambridge University Press

Csillag, F, Boots, B, Fortin, M, Lowell, K and Potvin, F 2001 'Multiscale characterization of boundaries and landscape ecological patterns'. *Geomatica* **55**(4), 509-522

Darby, H C 1956 *The Draining of the Fens*, 2nd edn. Cambridge: Cambridge University Press

Dormann, C F 2007 'Effects of incorporating spatial autocorrelation into the analysis of species distribution data'. *Global Ecology and Biogeography* **16**(2), 129-138

Dormann, C F, McPherson, J M, Araújo, M B, Bivand, R, Bolliger, J, Carl, G, Davies, R G, Hirzel, A, Jetz, W, Kissling, W D, Kühn, I, Ohlemüller, R, Peres-Neto, P R, Reineking, B, Schröder, B, Schurr, F M and Wilson, R 2007 'Methods to account for spatial autocorrelation in the analysis of species distributional data: a review'. *Ecography* **30**(5), 609–628

Dubayah, R and Rich, P M 1995 'Topographic solar radiation models for GIS'. *International Journal of Geographical Information Systems* **9**(4), 405–419

Dubin, R 2009 'Spatial Weights', in Fotheringham, A S and Rogerson, P A (eds) *The SAGE Handbook of Spatial Analysis*. London: SAGE Publications, 125–57

Duque, J C, Anselin, L and Rey, S J 2012 'The max-p-regions problem'. *Journal of Regional Science* **52**(3), 397-419

Duque, J C, Ramos, R and Suriñach, J 2007 'Supervised Regionalization Methods: A Survey'. *International Regional Science Review* **30**(3), 195-220

Dyer, C C 2003 'Review of B K Roberts and S Wrathmell, *Region and Place: A Study of English Rural Settlement*'. *Landscape History* **25**, 103–4

English Heritage 2011a *The National Heritage Protection Plan*. London: English Heritage. Retrieved 25 May 2011 from <http://www.english-heritage.org.uk/content/imported-docs/k-o/nhpp-plan.pdf>

English Heritage 2011b *The National Heritage Protection Plan. Measure 4. Understanding: Assessment of Character and Significance Activity Programme*. London: English Heritage. Retrieved 25 May 2011 from <http://www.english-heritage.org.uk/content/imported-docs/k-o/nhpp-measure4.pdf>

Esri 2011a *ArcGIS Desktop 10 Help: An introduction to interpolation methods* [web page]. Last updated 11 February 2011. Retrieved 15 February 2012 from <http://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#/An_introduction_to_interpolation_methods/003100000008000000/>

Esri 2011b *ArcGIS Desktop 10 Help: Area Solar Radiation (Spatial Analyst)* [web page]. Last updated 29 June 2011. Retrieved 27 October 2014 from <http://help.arcgis.com/en/arcgisdesktop/10.0/help/#/Area_Solar_Radiation/009z000000t5000000/>

Esri 2011c *ArcGIS Desktop 10 Help: How Iso Cluster works* [web page]. Last updated 29 June 2011. Retrieved 21 October 2014 from <http://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#/How_Iso_Cluster_works/009z000000q8000000/>

Esri 2011d *ArcGIS Desktop 10 Help: How solar radiation is calculated* [web page]. Last updated 29 June 2011. Retrieved 27 October 2014 from <http://help.arcgis.com/en/arcgisdesktop/10.0/help/#/How_solar_radiation_is_calculated/009z000000tm000000/>

Esri 2011e *ArcGIS Desktop 10 Help: Modeling solar radiation* [web page]. Last updated 29 June 2011. Retrieved 27 October 2014 from <http://help.arcgis.com/en/arcgisdesktop/10.0/help/#/Modeling_solar_radiation/009z000000t9000000/>

Esri 2012 *ArcGIS Desktop 10 Help: Classifying numerical fields for graduated symbology* [web page]. Last updated 8 February 2012. Retrieved 11 December 2014 from <http://help.arcgis.com/en/arcgisdesktop/10.0/help/#/Classifying_numerical_fields_for_graduated_symbology/00s50000001r000000/>

Esri 2013 *Interpreting OLS results* [web page]. Last updated 18 April 2013. Retrieved 12 February 2015 from <http://resources.arcgis.com/en/help/main/10.1/index.html#/Interpreting_OLS_results/005p00000030000000/>

Evans, I S 1977 'The selection of class intervals'. *Transactions of the Institute of British Geographers* new ser **2**(1), 98–124

Eve, S J and Crema, E R 2014 'A house with a view? Multi-model inference, visibility fields, and point process analysis of a Bronze Age settlement on Leskernick Hill (Cornwall, UK)'. *Journal of Archaeological Science* **43**(0), 267-277

Fortin, M-J and Dale, M R T 2005 *Spatial Analysis: A Guide for Ecologists*. Cambridge: Cambridge University Press

Fortin, M-J and Jacquez, G M 2000 'Randomization tests and spatially auto-correlated data'. *Bulletin of the Ecological Society of America* **81**(3), 201-205

Fortin, M J, Drapeau, P and Jacquez, G M 1996 'Quantification of the spatial co-occurrences of ecological boundaries'. *Oikos* **77**(1), 51-60

Fortin, M J, Keitt, T, Maurer, B, Taper, M, Kaufman, D M and Blackburn, T 2005 'Species' geographic ranges and distributional limits: pattern analysis and statistical issues'. *Oikos* **108**(1), 7-17

Fotheringham, A S, Brunsdon, C and Charlton, M 2002 *Geographically Weighted Regression: The analysis of spatially varying relationships*. Chichester: John Wiley & Sons

Freedman, D A 1983 'A Note on Screening Regression Equations'. *The American Statistician* **37**(2), 152-155

Fu, P and Rich, P M 2002 'A geometric solar radiation model with applications in agriculture and forestry'. *Computers and Electronics in Agriculture* **37**, 25–35

Gallo, K P, Easterling, D R and Peterson, T C 1996 'The Influence of Land Use/Land Cover on Climatological Values of the Diurnal Temperature Range'. *Journal of Climate* **9**(11), 2941–2944

Getis, A 2010 'Spatial autocorrelation', in Fischer, M M and Getis, A (eds) *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*. Berlin and Heidelberg: Springer-Verlag, 255–278

Gonçalves, M L, Netto, M L A, Costa, J A F and Zullo Júnior, J 2008 'An unsupervised method of classifying remotely sensed images using Kohonen self-organizing maps and agglomerative hierarchical clustering methods'. *International Journal of Remote Sensing* **29**(11), 3171–3207

Gregory, I and Ell, P S 2007 *Historical GIS: Technologies, Methodologies and Scholarship* (Cambridge Studies in Historical Geography **39**). Cambridge: Cambridge University Press

Gregory, I N 2008 '"A map is just a bad graph": Why spatial statistics are important in historical GIS', in Knowles, A K (ed) *Placing History: How Maps, Spatial Data, and GIS are Changing Historical Scholarship*. Redlands, CA: ESRI Press, 123–149

Grohmann, C H, Smith, M J and Riccomini, C 2011 'Multiscale Analysis of Topographic Surface Roughness in the Midland Valley, Scotland'. *IEEE Transactions on Geoscience and Remote Sensing* **49**(4), 1200–1213

Guo, D 2008 'Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP)'. *International Journal of Geographical Information Science* **22**(7), 801-823

Guo, D 2010 'Local entropy map: a nonparametric approach to detecting spatially varying multivariate relationships'. *International Journal of Geographical Information Science* **24**(9), 1367–1389

Guo, D, Chen, J, MacEachren, A M and Liao, K 2006 'A visualization system for space-time and multivariate patterns (vis-stamp)'. *Visualization and Computer Graphics, IEEE Transactions on* **12**(6), 1461-1474

Guo, D, Gahegan, M, MacEachren, A M and Zhou, B 2005 'Multivariate analysis and geovisualization with an integrated geographic knowledge discovery approach'. *Cartography and Geographic Information Science* **32**(2), 113-132

Guo, D and Wang, H 2011 'Automatic Region Building for Spatial Analysis'. *Transactions in GIS* **15**, 29–45

Haining, R P 2003 *Spatial Data Analysis: Theory and Practice*. Cambridge: Cambridge University Press

Hijmans, R J, Cameron, S E, Parra, J L, Jones, P G and Jarvis, A 2005 'Very high resolution interpolated climate surfaces for global land areas'. *International Journal of Climatology* **25**(15), 1965–1978

Hill, D 1998 'Eleventh century labours of the months in prose and pictures'. *Landscape History* **20**, 29–39

Hinton, D A 2005 'Debate: South Hampshire, "East Wessex" and the *Atlas of Rural Settlement in England*. *Landscape History* **27**, 71–5

Jacquez, G, Kaufmann, A and Goovaerts, P 2008 'Boundaries, links and clusters: a new paradigm in spatial analysis?'. *Environmental and Ecological Statistics* **15**(4), 403-419

Jacquez, G M 1995 'The map comparison problem: Tests for the overlap of geographic boundaries'. *Statistics in Medicine* **14**(21-22), 2343-2361

Jacquez, G M 2010 'Geographic boundary analysis in spatial and spatio-temporal epidemiology: Perspective and prospects'. *Spatial and Spatio-temporal Epidemiology* **1**(4), 207-218

Jain, A K 2010 'Data clustering: 50 years beyond K-means'. *Pattern Recognition Letters* **31**(8), 651-666

Jenks, G F 1977 *Optimal Data Classification for Choropleth Maps* (Department of Geography Occasional Paper **2**). Lawrence, Kansas: University of Kansas

Jenks, G F and Caspall, F C 1971 'Error on Choroplethic Maps: Definition, Measurement, Reduction'. *Annals of the Association of American Geographers* **61**(2), 217–244

Jenness, J S 2004 'Calculating landscape surface area from digital elevation models'. *Wildlife Society Bulletin* **32**(3), 829–839

Jones, R 2010 'The Village and the Butterfly: Nucleation out of Chaos and Complexity'. *Landscapes* **11**(1), 25–46

Kelejian, H H and Prucha, I R 2007 'HAC estimation in a spatial framework'. *Journal of Econometrics* **140**(1), 131-154

Kissling, W D and Carl, G 2007 'Spatial autocorrelation and the selection of simultaneous autoregressive models'. *Global Ecology and Biogeography* **17**(1), 59-71

Klippel, A, Hardisty, F and Li, R 2011 'Interpreting Spatial Patterns: An Inquiry Into Formal and Cognitive Aspects of Tobler's First Law of Geography'. *Annals of the Association of American Geographers* **101**(5), 1011-1031

Kohonen, T 2001 *Self-Organizing Maps*, 3rd edn (Springer Series in Information Sciences **30**). Berlin: Springer

Krzanowski, W J and Lai, Y T 1988 'A Criterion for Determining the Number of Groups in a Data Set Using Sum-of-Squares Clustering'. *Biometrics* **44**(1), 23-34

Lam, N S-N 1983 'Spatial Interpolation Methods: A Review'. *The American Cartographer* **10**(2), 129-150

Lambourne, A 2010 *Patterning within the Historic Landscape and its Possible Causes: A study of the incidence and origins of regional variation in southern England* (BAR Brit Ser **509**). Oxford: Archaeopress

Lawley, R 2009 *The Soil-Parent Material Database: a user guide*. (Open Report **OR/08/034**). Nottingham, UK: British Geological Survey, available from <http://nora.nerc.ac.uk/8048/>

Legendre, P 1993 'Spatial Autocorrelation: Trouble or New Paradigm?'. *Ecology* **74**(6), 1659–1673

Lillesand, T M, Kiefer, R W and Chipman, J W 2008 *Remote Sensing and Image Interpretation*, 6th edn. Hoboken, NJ: John Wiley & Sons

Ljungqvist, F C, Krusic, P J, Brattström, G and Sundqvist, H S 2012 'Northern Hemisphere temperature patterns in the last 12 centuries'. *Climate of the Past* **8**(1), 227–249

Lloyd, C D 2009 *Spatial Data Analysis: An Introduction for GIS Users*. Oxford: Oxford University Press

Lloyd, C D 2010 *Local Models for Spatial Analysis*, 2nd edn. Boca Raton, FL: CRC Press (Taylor & Francis Group)

Lo, A and MacKinlay, A 1990 'Data-snooping biases in tests of financial asset pricing models'. *Review of Financial Studies* **3**(3), 431-467

Lowerre, A 2010 'The Atlas of Rural Settlement in England GIS'. *Landscapes* **11**(2), 21–44

Lowerre, A 2011 'The Atlas of Rural Settlement in England GIS'. *Research News* **16**, 32–35

Lowerre, A G, Lyons, E, Roberts, B K and Wrathmell, S 2011 *The Atlas of Rural Settlement in England GIS: Data, Metadata and Documentation* [Computer file]. Swindon: English Heritage. Retrieved 23 March 2011 from <http://www.english-heritage.org.uk/professional/research/archaeology/atlas-of-rural-settlement-gis/>

MacEachren, A M 1995 *How Maps Work: Representation, Visualization, and Design*. New York and London: The Guilford Press

Manning, K, Downey, S S, Colledge, S, Conolly, J, Stopp, B, Dobney, K and Shennan, S 2013a 'The origins and spread of stock-keeping: the role of cultural and environmental influences on early Neolithic animal exploitation in Europe'. *Antiquity* **87**, 1046–1059

Manning, K, Stopp, B, Colledge, S, Downey, S, Conolly, J, Dobney, K and Shennan, S 2013b 'Animal exploitation in the early Neolithic of the Balkans and Central Europe', in Colledge, S, Conolly, J, Dobney, K, Manning, K and Shennan, S (eds) *The Origins and Spread of Domestic Animals in Southwest Asia and Europe* (Publications of the Institute of Archaeology, University College, London **59**). Walnut Creek, CA: Left Coast Press, 237–252

Maruca, S L and Jacquez, G M 2002 'Area-based tests for association between spatial patterns'. *Journal of Geographical Systems* **4**(1), 69-83

McCune, B and Keon, D 2002 'Equations for potential annual direct incident radiation and heat load'. *Journal of Vegetation Science* **13**(4), 603–606

Milligan, G and Cooper, M C 1985 'An examination of procedures for determining the number of clusters in a data set'. *Psychometrika* **50**(2), 159-179

Milligan, G and Cooper, M C 1988 'A study of standardization of variables in cluster analysis'. *Journal of Classification* **5**(2), 181-204

Mitchell, A 2005 *The ESRI Guide to GIS Analysis Volume 2: Spatial Measurements & Statistics*. Redlands, CA: ESRI Press

Muller, J-C 1975 'Associations in Choropleth Map Comparison'. *Annals of the Association of American Geographers* **65**(3), 403–413

Muller, J-C 1976 'Objective and Subjective Comparison in Choroplethic Mapping'. *The Cartographic Journal* **13**(2), 156-166

Muller, J-C 1979 'Perception of Continuously Shaded Maps'. *Annals of the Association of American Geographers* **69**(2), 240–249

National Soils Research Institute (NSRI) 2001 *NATMAP Soilscapes* [Computer file]. Cranfield: Cranfield University.

National Soils Research Institute (NSRI) 2013 *The Soils Guide* [web page]. Cranfield: Cranfield University. Retrieved 3 April 2013 from <www.landis.org.uk>

National Soils Research Institute (NSRI) 2014a *National Soil Map of England and Wales - NATMAP Vector* [web page]. Cranfield: Cranfield University. Retrieved 29 October 2014 from <http://www.landis.org.uk/data/nmvector.cfm>

National Soils Research Institute (NSRI) 2014b *National Soils Inventory - NSI* [web page]. Cranfield: Cranfield University. Retrieved 29 October 2014 from <http://www.landis.org.uk/data/nsi.cfm>

O'Brien, R M 2007 'A Caution Regarding Rules of Thumb for Variance Inflation Factors'. *Quality & Quantity* **41**(5), 673-690

Oden, N L, Sokal, R R, Fortin, M-J and Goebl, H 1993 'Categorical Wombling: Detecting Regions of Significant Change in Spatially Located Categorical Variables'. *Geographical Analysis* **25**(4), 315-336

Okabe, A, Boots, B, Sugihara, K and Chiu, S N 2000 *Spatial Tessellations: Concepts and Applications of Voronoi diagrams*, 2nd edn (Wiley Series in Probability and Statistics **501**). Chichester: Wiley

Oke, T R 1982 'The energetic basis of the urban heat island'. *Quarterly Journal of the Royal Meteorological Society* **108**(455), 1–24

Ordnance Survey 2010 *OS OpenData: Mapping data and geographic information from Ordnance Survey* [web page]. Retrieved 26 November 2010 from <http://www.ordnancesurvey.co.uk/oswebsite/opendata/>

Perry, M and Hollis, D 2005a 'The development of a new set of long-term climate averages for the UK'. *International Journal of Climatology* **25**(8), 1023–1039

Perry, M and Hollis, D 2005b 'The generation of monthly gridded datasets for a range of climatic variables over the UK'. *International Journal of Climatology* **25**(8), 1041–1054

Peterson, T C and Owen, T W 2005 'Urban Heat Island Assessment: Metadata Are Important'. *Journal of Climate* **18**(14), 2637–2646

Picard, R R and Berk, K N 1990 'Data Splitting'. *The American Statistician* **44**(2), 140-147

Riley, S J, DeGloria, S D and Elliot, R 1999 'A terrain ruggedness index that quantifies topographic heterogeneity'. *Intermountain Journal of Sciences* **5**(1-4), 23–27

Rippon, S 2010 'Landscape change during the 'Long Eighth Century' in southern England', in Higham, N J and Ryan, M J (eds) *The Landscape Archaeology of Anglo-Saxon England*. Woodbridge: Boydell & Brewer, 39–64

Rippon, S J, Fyfe, R M and Brown, A G 2006 'Beyond Villages and Open Fields: The Origins and Development of a Historic Landscape Characterised by Dispersed Settlement in South-West England'. *Medieval Archaeology* **50**, 31–70

Roberts, B K and Wrathmell, S 2000 *An Atlas of Rural Settlement in England*, 2003 corrected reprint edn. London: English Heritage

Sappington, J M, Longshore, K M and Thomson, D B 2007 'Quantifying Landscape Ruggedness for Animal Habitat Analysis: A case Study Using Bighorn Sheep in the Mojave Desert'. *Journal of Wildlife Management* **71**(5), 1419–1426

Sirami, C, Brotons, L and Martin, J-L 2009 'Do bird spatial distribution patterns reflect population trends in changing landscapes?'. *Landscape Ecology* **24**(7), 893-906

Sparks, P J and Sparks, C S 2009 'An Application of Spatially Autoregressive Models to the Study of US County Mortality Rates'. *Population, Space and Place* **16**(6), 465–481

Sugar, C A and James, G M 2003 'Finding the Number of Clusters in a Dataset: An Information-Theoretic Approach'. *Journal of the American Statistical Association* **98**(463), 750-763

Tibshirani, R, Walther, G and Hastie, T 2001 'Estimating the number of clusters in a data set via the gap statistic'. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63**(2), 411-423

Turner, S 2006 'Historic landscape characterisation: a landscape archaeology for research, management and planning'. *Landscape Research* 31(4), 385–398

Waller, L A and Jacquez, G M 1995 'Disease models implicit in statistical tests of disease clustering'. *Epidemiology* 6(6), 584-90

Wheeler, D and Calder, C 2007 'An assessment of coefficient accuracy in linear regression models with spatially varying coefficients'. *Journal of Geographical Systems* 9(2), 145–166

Wheeler, D and Tiefelsdorf, M 2005 'Multicollinearity and correlation among local regression coefficients in geographically weighted regression'. *Journal of Geographical Systems* 7(2), 161–187

Wheeler, D C 2007 'Diagnostic tools and a remedial method for collinearity in geographically weighted regression'. *Environment and Planning A* 39(10), 2464–2481

White, H 2000 'A Reality Check for Data Snooping'. *Econometrica* 68(5), 1097-1126

Whitson, J A and Galinsky, A D 2008 'Lacking Control Increases Illusory Pattern Perception'. *Science* 322(5898), 115-117

Williamson, T 2003 *Shaping Medieval Landscapes: Settlement, Society, Environment*. Macclesfield: Windgather Press

Williamson, T 2005 'Explaining regional landscapes: East Anglia and the Midlands in the Middle Ages', in Harper-Bill, C (ed) *Medieval East Anglia*. Woodbridge: Boydell, 11–32

Williamson, T 2007 'Historic landscape characterisation: some queries'. *Landscapes* 8(2), 64–71

Williamson, T 2010 'The environmental contexts of Anglo-Saxon settlement', in Higham, N J and Ryan, M J (eds) *The Landscape Archaeology of Anglo-Saxon England*. Woodbridge: Boydell & Brewer, 134–155

Williamson, T 2013 *Environment, Society and Landscape in Early Medieval England: Time and Topography* (Anglo-Saxon Studies 19). Woodbridge: The Boydell Press

Williamson, T, Liddiard, R and Partida, T 2013 *Champion: The Making and Unmaking of the English Midland Landscape*. Liverpool: Liverpool University Press

APPENDIX 1: GLOSSARY OF ABBREVIATIONS

ADS: Archaeology Data Service

AIC: Akaike's Information Criterion

AICc: Small-sample corrected version of Akaike's Information Criterion

BGS: British Geological Survey

CSS: Combined Settlement Score

CSR: Complete Spatial Randomness

DEM: Digital Elevation Model

GIS: Geographic Information System(s)

GWR: Geographically Weighted Regression

HLC: Historic Landscape Characterisation

NSRI: National Soils Research Institute

OLS: Ordinary Least-Squares (regression)

OS: Ordnance Survey

RAO: Relative Area Overlap

TRI: Topographic Roughness Index

VIF: Variance Inflation Factor

VRM: Vector Ruggedness Measure

APPENDIX 2: SOILSCAPE TYPES/TYPE COMBINATIONS AND CONSTITUENT SOIL ASSOCIATIONS

Soilscape ID	Combination ID	Description	Soil Associations
1	32	Saltmarsh soils	Saline 1
2	52	Shallow very acid peaty soils over rock	Bangor; Revidge; Skiddaw
3	54	Shallow lime-rich soils over chalk or limestone	Andover 1; Andover 2; Elmton 1; Elmton 2; Elmton 3; Icknield; Marcham; Newmarket 1; Newmarket 2; Reach; Sherborne; Upton 1; Upton 2; Wantage 1; Wantage 2; Wetton 1
4	32	Sand dune soils	Sandwich
5	57	Freely draining lime-rich loamy soils	Aberford; Aswarby; Badsey 1; Badsey 2; Blewbury; Block; Coombe 1; Coombe 2; Grove; Landbeach; Milton; Moreton; Panholes; Ruskington; Stretham; Swaffham Prior
6		Freely draining slightly acid loamy soils	Ardington; Barrow; Barton; Bearsted 1; Bearsted 2; Bromsgrove; Carstens; Charity 1; Crediton; Denbigh 1; Denbigh 2; Eardiston 1; Eardiston 2; East Keswick 1; East Keswick 2; Efford 1; Efford 2; Ellerbeck; Escrick 1; Escrick 2; Fyfield 1; Fyfield 2; Fyfield 3; Fyfield 4; Hamble 1; Hamble 2; Harwell; Hucklesbrook; Ludford; Marlow; Milford; Munslow; Neath; Newbiggin; Newnham; Oglethorpe; Rheidol; Rivington 1; Rivington 2; Rowton; Sonning 1; Sonning 2; South Petherton; Stone Street; Waterstock; Wick 1; Wick 2; Wick 3
7	54, 57	Freely draining slightly acid but base-rich soils	Banbury; Charity 2; Crwbin; East Keswick 3; Frilsham; Hunstanton; Malham 2; Malling; Melford; Moulton; Nordrach; Ston Easton; Sutton 1; Sutton 2; Tathwell; Trusham; Waltham
8		Slightly acid loamy and clayey soils with impeded drainage	Ashley; Batcombe; Bignor; Bishampton 1; Bishampton 2; Bromyard; Burlingham 1; Burlingham 2; Burlingham 3; Bursledon; Curtisden; Dunnington Heath; Flint; Halstow; Hodnet; Hornbeam 1; Hornbeam 2; Hornbeam 3; Middleton; Nercwys; Oxpasture; Ratsborough; Salwick; Stow; Tendring; Whimple 1; Whimple 2; Whimple 3; Wix; Worcester; Yeld
9		Lime-rich loamy and clayey soils with impeded drainage	Cannamore; Evesham 1; Evesham 2; Evesham 3; Hanslope
10	35	Freely draining slightly acid sandy soils	Bridgnorth; Cuckney 1; Cuckney 2; Downham; Frilford; Kexby; Newport 1; Newport 2; Newport 3; Newport 4; Ollerton
11	35	Freely draining sandy Breckland soils	Methwold; Worlington
12	33	Freely draining floodplain soils	Alun; Lugwardine; Teme; Wharfe
13		Freely draining acid loamy soils over rock	Dunwell; Malvern; Manod; Moor Gate; Moretonhampstead; Parc; Powys; Withnell 1; Withnell 2
14	35	Freely draining very acid sandy and loamy soils	Anglezarke; Crannymoor; Delamere; Goldstone; Larkbarrow; Shirrell Heath 1; Shirrell Heath 2; Southampton
15	35	Naturally wet very acid sandy and loamy soils	Blackwood; Bolderwood; Everingham; Felthorpe; Holidays Hill; Holme Moor; Poundgate; Sollom 1; Sollom 2
16	52	Very acid loamy upland soils with a wet peaty surface	Belmont; Earle; Gelligaer; Hafren; Hense; Hexworthy; Lydcott; Malham 1; Maw; Wetton 2
17		Slowly permeable seasonally wet acid loamy and clayey soils	Bardsey; Brickfield 1; Brickfield 2; Brickfield 3; Cegin; Claverley; Croft Pascoe; Dale; Dunkeswell; Essendon; Fforest; Gresham; Hallsworth 1; Hallsworth 2; Oak 1; Pinder; Sportsmans; Stanway; Vernolds
18		Slowly permeable seasonally wet slightly acid but base-rich loamy and clayey soils	Beccles 1; Beccles 2; Beccles 3; Brockhurst 1; Brockhurst 2; Clifton; Crewe; Denchworth; Dunkeswick; Foggathorpe 1; Foggathorpe 2; Holderness; Kingston; Martock; Oak 2; Ragdale; Rufford; Salop; Wickham 1; Wickham 2; Wickham 3; Wickham 4; Wickham 5; Windsor
19	52	Slowly permeable wet very acid upland soils with a peaty surface	Laployd; Onecote; Princetown; Wenallt; Wilcocks 1; Wilcocks 2
20	33	Loamy and clayey floodplain soils with naturally high groundwater	Compton; Conway; Enborne; Fladbury 1; Fladbury 2; Fladbury 3; Frome; Hollington; Midelney; Thames
21	32	Loamy and clayey soils of coastal flats with naturally high groundwater	Agney; Blacktoft; Dowels; Newchurch 1; Newchurch 2; Normoor; Rockcliffe; Romney; Tanvats; Wallasea 1; Wallasea 2; Wisbech
22	33	Loamy soils with naturally high groundwater	Arrow; Curdridge; Hurst; Kelmscot; Park Gate; Sessay; Shabbington; Swanwick; Wigton Moor; Yeollandpark
23	53	Loamy and sandy soils with naturally high groundwater and a peaty surface	Clayhythe; Downholland 1; Downholland 2; Downholland 3; Hanworth; Ireton; Isleham 1; Isleham 2; Peacock

Soilscape ID	Combination ID	Description	Soil Associations
24	53	Restored soils mostly from quarry and opencast spoil	Fly Restored Ironstone; Neutral Restored Opencast; Raw China Clay Spoil; Raw Slate Quarry Rubble; Restored Coprolite
25	52	Blanket bog peat soils	Crowdy 1; Crowdy 2; Winter Hill
26	53	Raised bog peat soils	Longmoss; Turbary Moor
27	53	Fen peat soils	Adventurers' 1; Adventurers' 2; Adventurers' 3; Altcar 1; Altcar 2; Mendham; Willingham

Combination ID	Constituent Soilscape IDs	Description
32	1, 4, 21	Coastal soils
33	12, 20, 22	Loamy floodplain soils and loamy soils with naturally high groundwater
35	10, 11, 14, 15	Sandy and very acid loamy soils (except for sand dunes [4])
52	2, 16, 19, 25	Upland peaty soils
53	23, 24, 26, 27	Lowland peaty soils plus restored soils
54	3, 7	Shallow lime-rich soils over chalk or limestone and freely draining slightly acid but base-rich soils
57	5, 7	Freely draining lime-rich loamy and slightly acid but base-rich soils

Soilscape descriptions and the lists of related soil Associations are derived from NSRI's 'The Soils Guide' web page (www.landis.org.uk).

APPENDIX 3: OLS BEST MODEL DETAILS

In the following tables, p-values in bold are significant at α = 0.01; p-values in italic are significant at α = 0.05.

Table 20: Regression results and diagnostics for DstNclAll Subset 2, model 1

Variable	Coefficient	Robust Standard Error	Robust t-statistic	Robust P-value	VIF
Intercept	7141.7	604.43	11.82	**0.0000**	-----
Elevation	3.0	0.36	8.15	**0.0000**	5.74
p_av2m45	-69.3	4.76	-14.56	**0.0000**	9.56
p_11	19.3	1.35	14.30	**0.0000**	6.62
t_8	-203.0	27.21	-7.46	**0.0000**	2.55
Soils Type 5	-733.9	135.55	-5.41	**0.0000**	3.26
Soils Type 6	-329.6	133.03	-2.48	*0.0132*	9.61
Soils Type 8	-195.4	130.63	-1.50	0.1347	7.15
Soils Type 9	-608.6	127.76	-4.76	**0.0000**	4.83
Soils Type 13	386.9	178.99	2.16	*0.0306*	3.12
Soils Type 17	-200.8	143.80	-1.40	0.1627	5.89
Soils Type 18	-317.7	129.71	-2.45	*0.0143*	12.98
Soils Combo 32	582.1	142.56	4.08	**0.0001**	4.95
Soils Combo 33	-969.2	161.47	-6.00	**0.0000**	2.27
Soils Combo 35	73.8	138.90	0.53	0.5951	5.04
Soils Combo 52	3246.2	247.06	13.14	**0.0000**	9.18
Soils Combo 53	897.3	167.90	5.34	**0.0000**	2.84
Soils Combo 54	-619.6	131.44	-4.71	**0.0000**	7.51

Response Variable	DstNclAll	Number of Observations	10,986
R^2	0.351	Adjusted R^2	0.350
Joint Wald statistic	2222.8	P-value	**0.0000**
Koenker (BP) statistic	1004.9	P-value	**0.0000**
Jarque-Bera statistic	408173.9	P-value	**0.0000**
Moran's I z-score	100.9	P-value	**0.0000**

Table 21: Regression results and diagnostics for DstNclAll Subset 2, model 2

Variable	Coefficient	Robust Standard Error	Robust t-statistic	Robust P-value	VIF
Intercept	7125.3	604.24	11.79	0.0000	-----
Elevation	3.0	0.36	8.26	0.0000	5.72
p_av2m45	-69.3	4.79	-14.45	0.0000	9.62
p_11	19.4	1.36	14.29	0.0000	6.61
t_8	-202.4	27.18	-7.45	0.0000	2.55
Soils Type 3	-626.8	133.20	-4.71	0.0000	5.82
Soils Type 6	-330.5	133.03	-2.48	*0.0130*	9.61
Soils Type 8	-194.8	130.65	-1.49	0.1359	7.15
Soils Type 9	-608.0	127.77	-4.76	0.0000	4.83
Soils Type 13	381.7	178.80	2.13	*0.0328*	3.12
Soils Type 17	-203.1	143.74	-1.41	0.1577	5.89
Soils Type 18	-316.9	129.73	-2.44	*0.0146*	12.98
Soils Combo 32	584.1	142.53	4.10	0.0000	4.95
Soils Combo 33	-968.6	161.48	-6.00	0.0000	2.27
Soils Combo 35	74.4	138.89	0.54	0.5922	5.04
Soils Combo 52	3237.8	245.96	13.16	0.0000	9.16
Soils Combo 53	899.5	167.80	5.36	0.0000	2.84
Soils Combo 57	-678.7	132.02	-5.14	0.0000	5.00

Response Variable	DstNclAll	Number of Observations	10,986
R^2	0.351	Adjusted R^2	0.350
Joint Wald statistic	2225.5	P-value	0.0000
Koenker (BP) statistic	1004.6	P-value	0.0000
Jarque-Bera statistic	408504.7	P-value	0.0000
Moran's I z-score	100.9	P-value	0.0000

Table 22: Regression results and diagnostics for DstNclAll Subset 2, model 3

Variable	Coefficient	Robust Standard Error	Robust t-statistic	Robust P-value	VIF
Intercept	7143.9	606.57	11.78	0.0000	-----
Elevation	3.0	0.36	8.16	0.0000	5.75
p_av2m45	-69.3	4.80	-14.45	0.0000	9.62
p_11	19.3	1.35	14.26	0.0000	6.63
t_8	-203.0	27.26	-7.45	0.0000	2.55
Soils Type 3	-623.1	133.25	-4.68	0.0000	5.82
Soils Type 5	-733.6	135.50	-5.41	0.0000	3.26
Soils Type 6	-329.6	133.03	-2.48	*0.0132*	9.61
Soils Type 7	-610.5	146.50	-4.17	0.0000	2.81
Soils Type 8	-195.4	130.64	-1.50	0.1347	7.15
Soils Type 9	-608.7	127.77	-4.76	0.0000	4.83
Soils Type 13	387.0	178.99	2.16	*0.0306*	3.12
Soils Type 17	-200.7	143.80	-1.40	0.1628	5.89
Soils Type 18	-317.8	129.73	-2.45	*0.0143*	12.98
Soils Combo 32	582.0	142.58	4.08	0.0001	4.95
Soils Combo 33	-969.3	161.49	-6.00	0.0000	2.27
Soils Combo 35	73.8	138.90	0.53	0.5950	5.04
Soils Combo 52	3246.4	247.16	13.13	0.0000	9.18
Soils Combo 53	897.3	167.90	5.34	0.0000	2.84

Response Variable	DstNclAll	Number of Observations	10,986
R^2	0.351	Adjusted R^2	0.350
Joint Wald statistic	2228.5	P-value	0.0000
Koenker (BP) statistic	1005.6	P-value	0.0000
Jarque-Bera statistic	408129.8	P-value	0.0000
Moran's I z-score	100.9	P-value	0.0000

Table 23: Regression results and diagnostics for DstNclBCD Subset 2, model 1

Variable	Coefficient	Robust Standard Error	Robust t-statistic	Robust P-value	VIF
Intercept	7942.6	1115.70	7.12	0.0000	-----
Elevation	5.5	0.45	12.27	0.0000	5.06
p_av2m34	-85.3	5.52	-15.44	0.0000	19.44
p_11	42.0	2.51	16.74	0.0000	13.77
DSR_av2m78	-9.8	2.26	-4.33	0.0000	2.13
Soils Type 3	-1118.2	169.38	-6.60	0.0000	5.84
Soils Type 6	-370.8	170.52	-2.17	*0.0297*	9.62
Soils Type 8	-176.2	167.05	-1.05	0.2915	7.15
Soils Type 9	-848.1	162.09	-5.23	0.0000	4.82
Soils Type 13	657.2	233.92	2.81	0.0050	3.18
Soils Type 17	-89.8	180.02	-0.50	0.6177	5.75
Soils Type 18	-418.6	163.25	-2.56	*0.0104*	12.88
Soils Combo 32	751.5	178.55	4.21	0.0000	4.94
Soils Combo 33	-1105.6	201.46	-5.49	0.0000	2.23
Soils Combo 35	122.5	174.01	0.70	0.4816	5.00
Soils Combo 52	3396.0	303.17	11.20	0.0000	8.99
Soils Combo 53	964.2	198.69	4.85	0.0000	2.81
Soils Combo 57	-1015.9	167.94	-6.05	0.0000	4.98

Response Variable	DstNclBCD	Number of Observations	10,986
R^2	0.334	Adjusted R^2	0.333
Joint Wald statistic	2424.2	P-value	0.0000
Koenker (BP) statistic	1069.2	P-value	0.0000
Jarque-Bera statistic	140527.4	P-value	0.0000
Moran's I z-score	106.9	P-value	0.0000

Table 24: Regression results and diagnostics for DstNclBCD Subset 2, model 2

Variable	Coefficient	Robust Standard Error	Robust t-statistic	Robust P-value	VIF
Intercept	7781.9	1056.90	7.36	0.0000	-----
Elevation	5.5	0.44	12.44	0.0000	5.03
p_av2m34	-85.3	5.52	-15.46	0.0000	19.45
p_11	42.0	2.51	16.72	0.0000	13.78
DSR_7	-8.9	2.02	-4.42	0.0000	2.07
Soils Type 3	-1120.1	169.40	-6.61	0.0000	5.84
Soils Type 6	-370.4	170.54	-2.17	*0.0299*	9.62
Soils Type 8	-174.9	167.05	-1.05	0.2953	7.15
Soils Type 9	-846.9	162.09	-5.22	0.0000	4.82
Soils Type 13	666.7	233.20	2.86	0.0043	3.17
Soils Type 17	-81.7	180.19	-0.45	0.6502	5.76
Soils Type 18	-415.8	163.27	-2.55	*0.0109*	12.88
Soils Combo 32	754.4	178.57	4.23	0.0000	4.94
Soils Combo 33	-1101.3	201.48	-5.47	0.0000	2.23
Soils Combo 35	127.2	174.01	0.73	0.4649	5.00
Soils Combo 52	3404.0	303.12	11.23	0.0000	9.00
Soils Combo 53	965.7	198.65	4.86	0.0000	2.81
Soils Combo 57	-1013.2	167.95	-6.03	0.0000	4.98

Response Variable	DstNclBCD	Number of Observations	10,986
R^2	0.334	Adjusted R^2	0.333
Joint Wald statistic	2432.8	P-value	0.0000
Koenker (BP) statistic	1063.6	P-value	0.0000
Jarque-Bera statistic	141141.2	P-value	0.0000
Moran's I z-score	106.8	P-value	0.0000

Table 25: Regression results and diagnostics for DstNclBCD Subset 2, model 3

Variable	Coefficient	Robust Standard Error	Robust t-statistic	Robust P-value	VIF
Intercept	7888.6	1113.41	7.09	0.0000	-----
Elevation	5.5	0.45	12.18	0.0000	5.05
p_av2m34	-84.7	5.45	-15.54	0.0000	19.11
p_11	41.7	2.48	16.81	0.0000	13.62
DSR_av2m78	-9.7	2.26	-4.29	0.0000	2.13
Soils Type 5	-1073.3	171.00	-6.28	0.0000	3.25
Soils Type 6	-369.1	170.53	-2.16	*0.0304*	9.62
Soils Type 8	-175.2	167.02	-1.05	0.2941	7.15
Soils Type 9	-847.4	162.06	-5.23	0.0000	4.82
Soils Type 13	663.5	234.36	2.83	0.0047	3.19
Soils Type 17	-88.4	180.16	-0.49	0.6236	5.75
Soils Type 18	-417.3	163.22	-2.56	*0.0106*	12.88
Soils Combo 32	751.0	178.57	4.21	0.0000	4.94
Soils Combo 33	-1103.7	201.41	-5.48	0.0000	2.23
Soils Combo 35	122.5	173.99	0.70	0.4814	5.00
Soils Combo 52	3400.2	304.47	11.17	0.0000	9.02
Soils Combo 53	962.7	198.74	4.84	0.0000	2.81
Soils Combo 54	-1067.7	167.72	-6.37	0.0000	7.52

Response Variable	DstNclBCD	Number of Observations	10,986
R^2	0.334	Adjusted R^2	0.333
Joint Wald statistic	2439.4	P-value	0.0000
Koenker (BP) statistic	1067.7	P-value	0.0000
Jarque-Bera statistic	140623.8	P-value	0.0000
Moran's I z-score	106.9	P-value	0.0000

Table 26: Regression results and diagnostics for DstNclBCD Subset 3, model 1

Variable	Coefficient	Robust Standard Error	Robust t-statistic	Robust P-value	VIF
Intercept	6682.5	638.44	10.47	0.0000	-----
Elevation	6.1	0.46	13.19	0.0000	6.05
p_av3m345	-88.5	5.47	-16.18	0.0000	16.45
p_11	36.3	2.21	16.37	0.0000	11.58
t_8	-168.1	30.46	-5.52	0.0000	2.53
Soils Type 3	-999.4	170.21	-5.87	0.0000	5.71
Soils Type 6	-257.5	172.01	-1.50	0.1344	9.36
Soils Type 8	-165.4	169.17	-0.98	0.3281	7.07
Soils Type 9	-885.7	164.09	-5.40	0.0000	4.60
Soils Type 13	1039.7	247.40	4.20	0.0000	2.93
Soils Type 17	-299.4	188.33	-1.59	0.1120	5.72
Soils Type 18	-440.1	167.68	-2.62	0.0087	12.56
Soils Combo 32	728.9	181.51	4.02	0.0001	4.82
Soils Combo 33	-946.4	204.73	-4.62	0.0000	2.18
Soils Combo 35	89.0	175.80	0.51	0.6126	5.00
Soils Combo 52	2829.2	293.05	9.65	0.0000	8.98
Soils Combo 53	887.5	197.66	4.49	0.0000	2.89
Soils Combo 57	-1181.8	169.26	-6.98	0.0000	4.96

Response Variable	DstNclBCD	Number of Observations	10,986
R^2	0.337	Adjusted R^2	0.336
Joint Wald statistic	2600.7	P-value	0.0000
Koenker (BP) statistic	1099.2	P-value	0.0000
Jarque-Bera statistic	154871.3	P-value	0.0000
Moran's I z-score	108.5	P-value	0.0000

Table 27: Regression results and diagnostics for DstNclBCD Subset 3, model 2

Variable	Coefficient	Robust Standard Error	Robust t-statistic	Robust P-value	VIF
Intercept	6659.1	641.24	10.38	0.0000	-----
Elevation	6.1	0.47	13.15	0.0000	6.07
p_av3m345	-88.3	5.48	-16.11	0.0000	16.48
p_11	36.3	2.21	16.38	0.0000	11.58
t_8	-167.4	30.54	-5.48	0.0000	2.54
Soils Type 3	-1002.1	170.30	-5.88	0.0000	5.71
Soils Type 5	-1120.8	173.06	-6.48	0.0000	3.31
Soils Type 6	-257.4	172.01	-1.50	0.1346	9.36
Soils Type 7	-1260.8	183.12	-6.89	0.0000	2.72
Soils Type 8	-164.5	169.16	-0.97	0.3308	7.07
Soils Type 9	-884.6	164.09	-5.39	0.0000	4.60
Soils Type 13	1034.3	247.67	4.18	0.0000	2.93
Soils Type 17	-301.3	188.40	-1.60	0.1098	5.72
Soils Type 18	-438.3	167.70	-2.61	0.0090	12.56
Soils Combo 32	732.1	181.61	4.03	0.0001	4.82
Soils Combo 33	-943.7	204.78	-4.61	0.0000	2.18
Soils Combo 35	90.1	175.81	0.51	0.6083	5.00
Soils Combo 52	2820.7	294.27	9.59	0.0000	9.00
Soils Combo 53	890.5	197.77	4.50	0.0000	2.89

Response Variable	DstNclBCD	Number of Observations	10,986
R^2	0.337	Adjusted R^2	0.336
Joint Wald statistic	2608.8	P-value	0.0000
Koenker (BP) statistic	1099.5	P-value	0.0000
Jarque-Bera statistic	155054.5	P-value	0.0000
Moran's I z-score	108.5	P-value	0.0000

Table 28: Regression results and diagnostics for DstNclBCD Subset 3, model 3

Variable	Coefficient	Robust Standard Error	Robust t-statistic	Robust P-value	VIF
Intercept	6697.6	639.67	10.47	0.0000	-----
Elevation	6.2	0.47	13.20	0.0000	6.06
p_av3m345	-89.3	5.43	-16.45	0.0000	16.30
p_11	36.6	2.20	16.62	0.0000	11.51
t_8	-168.2	30.51	-5.51	0.0000	2.53
Soils Type 5	-1114.8	173.05	-6.44	0.0000	3.31
Soils Type 6	-257.3	172.03	-1.50	0.1348	9.36
Soils Type 8	-163.5	169.17	-0.97	0.3337	7.07
Soils Type 9	-884.8	164.10	-5.39	0.0000	4.60
Soils Type 13	1036.4	247.66	4.18	0.0000	2.93
Soils Type 17	-299.6	188.41	-1.59	0.1119	5.72
Soils Type 18	-439.7	167.70	-2.62	0.0087	12.56
Soils Combo 32	730.6	181.60	4.02	0.0001	4.82
Soils Combo 33	-947.8	204.77	-4.63	0.0000	2.18
Soils Combo 35	91.2	175.83	0.52	0.6041	5.00
Soils Combo 52	2824.6	294.17	9.60	0.0000	9.00
Soils Combo 53	891.2	197.75	4.51	0.0000	2.89
Soils Combo 54	-1071.2	168.66	-6.35	0.0000	7.33

Response Variable	DstNclBCD	Number of Observations	10,986
R^2	0.337	Adjusted R^2	0.336
Joint Wald statistic	2594.0	P-value	0.0000
Koenker (BP) statistic	1099.2	P-value	0.0000
Jarque-Bera statistic	154657.2	P-value	0.0000
Moran's I z-score	108.5	P-value	0.0000

Table 29: Regression results and diagnostics for CSS Na2 Subset 2, model 1

Variable	Coefficient	Robust Standard Error	Robust t-statistic	Robust P-value	VIF
Intercept	1.494	0.0197	75.85	0.0000	-----
Elevation	< -0.001	0.0000	-15.43	0.0000	7.97
p_3	0.003	0.0002	15.43	0.0000	15.35
p_9	-0.004	0.0002	-23.91	0.0000	10.21
bio1	-0.052	0.0018	-29.66	0.0000	4.49
Soils Type 3	0.070	0.0063	11.20	0.0000	5.77
Soils Type 5	0.048	0.0069	6.99	0.0000	3.26
Soils Type 6	-0.032	0.0066	-4.86	0.0000	9.52
Soils Type 7	0.021	0.0075	2.76	0.0057	2.82
Soils Type 8	-0.052	0.0066	-7.90	0.0000	7.14
Soils Type 9	0.007	0.0067	1.05	0.2915	4.83
Soils Type 13	-0.042	0.0094	-4.49	0.0000	3.06
Soils Type 17	-0.081	0.0073	-11.19	0.0000	5.86
Soils Type 18	-0.028	0.0064	-4.44	0.0000	12.96
Soils Combo 32	-0.037	0.0073	-5.01	0.0000	4.97
Soils Combo 33	0.028	0.0088	3.14	0.0017	2.26
Soils Combo 35	-0.021	0.0068	-3.08	0.0021	5.04
Soils Combo 52	-0.074	0.0088	-8.40	0.0000	9.15
Soils Combo 53	-0.007	0.0075	-0.93	0.3513	2.85

Response Variable	CSS Na2	Number of Observations	10,986
R^2	0.282	Adjusted R^2	0.281
Joint Wald statistic	6263.3	P-value	0.0000
Koenker (BP) statistic	445.4	P-value	0.0000
Jarque-Bera statistic	3699.9	P-value	0.0000
Moran's I z-score	153.7	P-value	0.0000

Table 30: Regression results and diagnostics for CSS Nb2 Subset 2, model 1

Variable	Coefficient	Robust Standard Error	Robust t-statistic	Robust P-value	VIF
Intercept	1.512	0.0199	76.18	0.0000	-----
Elevation	< -0.001	0.0000	-16.74	0.0000	7.97
p_3	0.003	0.0002	15.80	0.0000	15.35
p_9	-0.004	0.0002	-24.41	0.0000	10.21
bio1	-0.054	0.0018	-30.53	0.0000	4.49
Soils Type 3	0.075	0.0065	11.46	0.0000	5.77
Soils Type 5	0.052	0.0071	7.30	0.0000	3.26
Soils Type 6	-0.033	0.0068	-4.81	0.0000	9.52
Soils Type 7	0.025	0.0078	3.25	0.0012	2.82
Soils Type 8	-0.053	0.0068	-7.82	0.0000	7.14
Soils Type 9	0.010	0.0070	1.45	0.1478	4.83
Soils Type 13	-0.048	0.0096	-4.97	0.0000	3.06
Soils Type 17	-0.082	0.0075	-10.94	0.0000	5.86
Soils Type 18	-0.027	0.0066	-4.16	0.0000	12.96
Soils Combo 32	-0.039	0.0076	-5.11	0.0000	4.97
Soils Combo 33	0.028	0.0090	3.13	0.0017	2.26
Soils Combo 35	-0.022	0.0071	-3.12	0.0018	5.04
Soils Combo 52	-0.072	0.0090	-7.94	0.0000	9.15
Soils Combo 53	-0.007	0.0076	-0.94	0.3447	2.85

Response Variable	CSS Nb2	Number of Observations	10,986
R^2	0.292	Adjusted R^2	0.290
Joint Wald statistic	6599.1	P-value	0.0000
Koenker (BP) statistic	430.3	P-value	0.0000
Jarque-Bera statistic	3052.6	P-value	0.0000
Moran's I z-score	152.8	P-value	0.0000

APPENDIX 4: SPATIAL REGRESSION MODEL DETAILS

In the following tables, p-values in bold are significant at $\alpha = 0.01$; p-values in italic are significant at $\alpha = 0.05$.

Table 31: Spatial lag regression results and diagnostics for DstNclAll Subset 2, model 1, spatial weights matrix 1

Variable	Coefficient	Std Error	z-value	Probability
W_DstNclAll	0.80	0.01	118.68	**0.0000**
Intercept	818.35	292.41	2.80	**0.0051**
Elevation	1.10	0.21	5.20	**0.0000**
p_av2m45	-13.79	3.01	-4.58	**0.0000**
p_11	3.48	1.02	3.40	**0.0007**
t_8	-6.06	14.41	-0.42	0.6738
Soils Type 5	-228.09	114.91	-1.99	*0.0472*
Soils Type 6	-138.60	102.33	-1.35	0.1756
Soils Type 8	17.59	101.90	0.17	0.8630
Soils Type 9	-88.94	106.06	-0.84	0.4017
Soils Type 13	124.95	127.12	0.98	0.3256
Soils Type 17	-102.93	107.71	-0.96	0.3393
Soils Type 18	6.83	100.40	0.07	0.9458
Soils Combo 32	358.94	107.08	3.35	**0.0008**
Soils Combo 33	-515.74	141.37	-3.65	**0.0003**
Soils Combo 35	122.65	107.27	1.14	0.2529
Soils Combo 52	982.34	123.31	7.97	**0.0000**
Soils Combo 53	394.57	118.17	3.34	**0.0008**
Soils Combo 54	-82.65	102.83	-0.80	0.4215

Response Variable	DstNclAll
Num of Observations	10,986
Pseudo-R^2	0.720
Breusch-Pagan statistic	948.6
P-value	**0.0000**

Table 32: Spatial error regression results and diagnostics for DstNclAll Subset 2, model 1, spatial weights matrix 1

Variable	Coefficient	Std Error	z-value	Probability
Intercept	6519.24	1351.76	4.82	**0.0000**
Elevation	2.77	0.60	4.62	**0.0000**
p_av2m45	-46.32	13.74	-3.37	**0.0008**
p_11	25.25	5.07	4.98	**0.0000**
t_8	-272.19	74.06	-3.68	**0.0002**
Soils Type 5	-394.69	152.77	-2.58	**0.0098**
Soils Type 6	-230.71	129.62	-1.78	0.0751
Soils Type 8	-22.54	132.99	-0.17	0.8654
Soils Type 9	-108.05	148.13	-0.73	0.4658
Soils Type 13	172.48	168.33	1.03	0.3055
Soils Type 17	-65.10	141.97	-0.46	0.6466
Soils Type 18	86.66	126.94	0.68	0.4948
Soils Combo 32	838.44	151.49	5.54	**0.0000**
Soils Combo 33	-410.18	165.68	-2.48	*0.0133*
Soils Combo 35	153.87	138.20	1.11	0.2656
Soils Combo 52	805.10	161.31	4.99	**0.0000**
Soils Combo 53	439.63	160.87	2.73	**0.0063**
Soils Combo 54	-131.89	135.80	-0.97	0.3314
Lambda	0.84	0.01	129.72	**0.0000**

Response Variable	DstNclAll
Num of Observations	10,986
Pseudo-R^2	0.721
Breusch-Pagan statistic	1108.2
P-value	**0.0000**

Table 33: Spatial lag regression results and diagnostics for DstNclAll Subset 2, model 2, spatial weights matrix 1

Variable	Coefficient	Std Error	z-value	Probability
W_DstNclAll	0.80	0.01	118.76	**0.000**
Intercept	784.26	292.36	2.68	**0.007**
Elevation	1.10	0.21	5.19	**0.000**
p_av2m45	-13.35	3.02	-4.42	**0.000**
p_11	3.44	1.02	3.37	**0.001**
t_8	-5.17	14.40	-0.36	0.720
Soils Type 3	-45.79	105.30	-0.44	0.664
Soils Type 6	-139.01	102.31	-1.36	0.174
Soils Type 8	18.19	101.88	0.18	0.858
Soils Type 9	-87.75	106.04	-0.83	0.408
Soils Type 13	122.11	127.03	0.96	0.336
Soils Type 17	-104.71	107.68	-0.97	0.331
Soils Type 18	8.31	100.38	0.08	0.934
Soils Combo 32	360.74	107.05	3.37	**0.001**
Soils Combo 33	-514.21	141.35	-3.64	**0.000**
Soils Combo 35	122.66	107.25	1.14	0.253
Soils Combo 52	976.21	123.11	7.93	**0.000**
Soils Combo 53	395.66	118.14	3.35	**0.001**
Soils Combo 57	-209.57	106.61	-1.97	*0.049*

Response Variable	DstNclAll
Num of Observations	10,986
Pseudo-R^2	0.720
Breusch-Pagan statistic	950.3
P-value	**0.0000**

Table 34: Spatial error regression results and diagnostics for DstNclAll Subset 2, model 2, spatial weights matrix 1

Variable	Coefficient	Std Error	z-value	Probability
Intercept	6474.78	1352.94	4.79	**0.0000**
Elevation	2.72	0.60	4.52	**0.0000**
p_av2m45	-45.52	13.75	-3.31	**0.0009**
p_11	25.33	5.08	4.99	**0.0000**
t_8	-272.54	74.13	-3.68	**0.0002**
Soils Type 3	-40.61	143.67	-0.28	0.7774
Soils Type 6	-230.68	129.60	-1.78	0.0751
Soils Type 8	-18.14	132.96	-0.14	0.8915
Soils Type 9	-107.02	148.10	-0.72	0.4699
Soils Type 13	164.11	168.29	0.98	0.3295
Soils Type 17	-68.58	141.95	-0.48	0.6290
Soils Type 18	94.53	126.93	0.75	0.4564
Soils Combo 32	840.46	151.47	5.55	**0.0000**
Soils Combo 33	-401.27	165.65	-2.42	*0.0154*
Soils Combo 35	161.98	138.18	1.17	0.2411
Soils Combo 52	801.42	161.23	4.97	**0.0000**
Soils Combo 53	448.84	160.85	2.79	**0.0053**
Soils Combo 57	-319.00	138.63	-2.30	*0.0214*
Lambda	0.84	0.01	129.86	**0.0000**

Response Variable	DstNclAll
Num of Observations	10,986
Pseudo-R^2	0.722
Breusch-Pagan statistic	1108.9
P-value	**0.0000**

Table 35: Spatial lag regression results and diagnostics for DstNclAll Subset 2, model 3, spatial weights matrix 1

Variable	Coefficient	Std Error	z-value	Probability
W_DstNclAll	0.80	0.01	118.75	**0.0000**
Intercept	791.80	292.75	2.71	**0.0068**
Elevation	1.09	0.21	5.15	**0.0000**
p_av2m45	-13.37	3.02	-4.43	**0.0000**
p_11	3.41	1.02	3.34	**0.0008**
t_8	-5.42	14.41	-0.38	0.7070
Soils Type 3	-44.36	105.34	-0.42	0.6737
Soils Type 5	-231.34	114.91	-2.01	*0.0441*
Soils Type 6	-138.68	102.31	-1.36	0.1753
Soils Type 7	-182.54	119.14	-1.53	0.1255
Soils Type 8	17.95	101.88	0.18	0.8602
Soils Type 9	-88.04	106.04	-0.83	0.4064
Soils Type 13	124.22	127.10	0.98	0.3284
Soils Type 17	-103.77	107.69	-0.96	0.3353
Soils Type 18	7.96	100.38	0.08	0.9368
Soils Combo 32	359.93	107.06	3.36	**0.0008**
Soils Combo 33	-514.49	141.35	-3.64	**0.0003**
Soils Combo 35	122.44	107.25	1.14	0.2536
Soils Combo 52	979.65	123.30	7.95	**0.0000**
Soils Combo 53	394.79	118.15	3.34	**0.0008**

Response Variable	DstNclAll
Num of Observations	10,986
Pseudo-R^2	0.720
Breusch-Pagan statistic	950.5
P-value	**0.0000**

Table 36: Spatial error regression results and diagnostics for DstNclAll Subset 2, model 3, spatial weights matrix 1

Variable	Coefficient	Std Error	z-value	Probability
Intercept	6488.46	1353.09	4.80	**0.0000**
Elevation	2.71	0.60	4.51	**0.0000**
p_av2m45	-45.56	13.75	-3.31	**0.0009**
p_11	25.25	5.08	4.97	**0.0000**
t_8	-272.72	74.13	-3.68	**0.0002**
Soils Type 3	-46.03	143.77	-0.32	0.7488
Soils Type 5	-383.48	152.86	-2.51	*0.0121*
Soils Type 6	-230.40	129.59	-1.78	0.0754
Soils Type 7	-256.20	152.15	-1.68	0.0922
Soils Type 8	-20.21	132.97	-0.15	0.8792
Soils Type 9	-108.87	148.11	-0.74	0.4623
Soils Type 13	167.87	168.32	1.00	0.3186
Soils Type 17	-67.51	141.95	-0.48	0.6344
Soils Type 18	91.94	126.95	0.72	0.4689
Soils Combo 32	839.30	151.47	5.54	**0.0000**
Soils Combo 33	-404.78	165.67	-2.44	*0.0146*
Soils Combo 35	158.90	138.20	1.15	0.2503
Soils Combo 52	805.74	161.28	5.00	**0.0000**
Soils Combo 53	445.44	160.88	2.77	**0.0056**
Lambda	0.84	0.01	129.88	**0.0000**

Response Variable	DstNclAll
Num of Observations	10,986
Pseudo-R^2	0.722
Breusch-Pagan statistic	1109.2
P-value	**0.0000**

Table 37: Spatial lag regression results and diagnostics for DstNclBCD Subset 2, model 1, spatial weights matrix 1

Variable	Coefficient	Std Error	z-value	Probability
W_DstNclBCD	0.83	0.01	137.05	**0.0000**
Intercept	1976.78	532.51	3.71	**0.0002**
Elevation	1.37	0.23	6.06	**0.0000**
p_av2m34	-16.46	3.24	-5.08	**0.0000**
p_11	7.06	1.66	4.26	**0.0000**
DSR_av2m78	-2.98	1.07	-2.79	**0.0053**
Soils Type 3	-126.36	118.48	-1.07	0.2862
Soils Type 6	-155.41	114.90	-1.35	0.1762
Soils Type 8	21.39	114.35	0.19	0.8516
Soils Type 9	-118.60	118.94	-1.00	0.3187
Soils Type 13	177.27	143.97	1.23	0.2182
Soils Type 17	-78.16	119.47	-0.65	0.5130
Soils Type 18	5.89	112.22	0.05	0.9581
Soils Combo 32	437.50	120.10	3.64	**0.0003**
Soils Combo 33	-507.91	157.25	-3.23	**0.0012**
Soils Combo 35	154.30	119.93	1.29	0.1982
Soils Combo 52	889.18	136.89	6.50	**0.0000**
Soils Combo 53	380.40	132.07	2.88	**0.0040**
Soils Combo 57	-255.44	119.49	-2.14	*0.0325*

Response Variable	DstNclBCD
Num of Observations	10,986
Pseudo-R^2	0.743
Breusch-Pagan statistic	570.0
P-value	**0.0000**

Table 38: Spatial error regression results and diagnostics for DstNclBCD Subset 2, model 1, spatial weights matrix 1

Variable	Coefficient	Std Error	z-value	Probability
Intercept	6321.36	826.28	7.65	**0.0000**
Elevation	4.15	0.61	6.87	**0.0000**
p_av2m34	-37.79	15.42	-2.45	*0.0142*
p_11	36.01	9.08	3.97	**0.0001**
DSR_av2m78	-10.99	1.62	-6.80	**0.0000**
Soils Type 3	-136.20	162.19	-0.84	0.4011
Soils Type 6	-229.50	145.57	-1.58	0.1149
Soils Type 8	10.66	149.32	0.07	0.9431
Soils Type 9	-122.94	166.58	-0.74	0.4605
Soils Type 13	117.43	191.18	0.61	0.5391
Soils Type 17	48.08	159.37	0.30	0.7629
Soils Type 18	184.75	142.38	1.30	0.1944
Soils Combo 32	1076.98	170.64	6.31	**0.0000**
Soils Combo 33	-250.88	185.51	-1.35	0.1762
Soils Combo 35	300.93	155.14	1.94	0.0524
Soils Combo 52	805.77	181.30	4.44	**0.0000**
Soils Combo 53	512.26	180.72	2.84	**0.0046**
Soils Combo 57	-324.45	155.78	-2.08	*0.0373*
Lambda	0.87	0.01	150.93	**0.0000**

Response Variable	DstNclBCD
Num of Observations	10,986
Pseudo-R^2	0.747
Breusch-Pagan statistic	643.4
P-value	**0.0000**

Table 39: Spatial lag regression results and diagnostics for DstNclBCD Subset 2, model 2, spatial weights matrix 1

Variable	Coefficient	Std Error	z-value	Probability
W_DstNclBCD	0.84	0.01	137.05	**0.0000**
Intercept	2073.77	516.54	4.02	**0.0001**
Elevation	1.37	0.23	6.08	**0.0000**
p_av2m34	-16.54	3.24	-5.11	**0.0000**
p_11	7.03	1.66	4.23	**0.0000**
DSR_7	-3.01	0.98	-3.08	**0.0021**
Soils Type 3	-128.97	118.48	-1.09	0.2764
Soils Type 6	-157.27	114.89	-1.37	0.1710
Soils Type 8	20.23	114.34	0.18	0.8595
Soils Type 9	-118.76	118.93	-1.00	0.3180
Soils Type 13	173.00	143.80	1.20	0.2290
Soils Type 17	-73.73	119.51	-0.62	0.5373
Soils Type 18	7.85	112.23	0.07	0.9443
Soils Combo 32	439.92	120.11	3.66	**0.0003**
Soils Combo 33	-506.77	157.24	-3.22	**0.0013**
Soils Combo 35	156.51	119.93	1.31	0.1919
Soils Combo 52	895.26	136.95	6.54	**0.0000**
Soils Combo 53	382.84	132.07	2.90	**0.0038**
Soils Combo 57	-255.14	119.48	-2.14	*0.0327*

Response Variable	DstNclBCD
Num of Observations	10,986
Pseudo-R^2	0.743
Breusch-Pagan statistic	557.9
P-value	**0.0000**

Table 40: Spatial error regression results and diagnostics for DstNclBCD Subset 2, model 2, spatial weights matrix 1

Variable	Coefficient	Std Error	z-value	Probability
Intercept	6092.85	799.93	7.62	**0.0000**
Elevation	4.07	0.61	6.72	**0.0000**
p_av2m34	-35.22	15.48	-2.28	*0.0228*
p_11	35.08	9.11	3.85	**0.0001**
DSR_7	-10.08	1.49	-6.76	**0.0000**
Soils Type 3	-129.97	162.13	-0.80	0.4228
Soils Type 6	-225.32	145.53	-1.55	0.1216
Soils Type 8	15.27	149.29	0.10	0.9185
Soils Type 9	-118.95	166.57	-0.71	0.4752
Soils Type 13	125.31	191.02	0.66	0.5118
Soils Type 17	49.62	159.36	0.31	0.7555
Soils Type 18	187.11	142.37	1.31	0.1887
Soils Combo 32	1078.26	170.64	6.32	**0.0000**
Soils Combo 33	-250.06	185.49	-1.35	0.1776
Soils Combo 35	305.94	155.12	1.97	*0.0486*
Soils Combo 52	802.50	181.28	4.43	**0.0000**
Soils Combo 53	512.66	180.71	2.84	**0.0046**
Soils Combo 57	-320.29	155.75	-2.06	*0.0397*
Lambda	0.87	0.01	151.19	**0.0000**

Response Variable	DstNclBCD
Num of Observations	10,986
Pseudo-R^2	0.747
Breusch-Pagan statistic	631.5
P-value	**0.0000**
Moran's I statistic	-0.0351

Table 41: Spatial lag regression results and diagnostics for DstNclBCD Subset 2, model 3, spatial weights matrix 1

Variable	Coefficient	Std Error	z-value	Probability		
W_DstNclBCD	0.84	0.01	137.00	**0.0000**	Response Variable	DstNclBCD
Intercept	2002.85	532.14	3.76	**0.0002**	Num of Observations	10,986
Elevation	1.38	0.23	6.13	**0.0000**	Pseudo-R^2	0.743
p_av2m34	-17.08	3.21	-5.32	**0.0000**	Breusch-Pagan statistic	568.9
p_11	7.29	1.65	4.42	**0.0000**	P-value	**0.0000**
DSR_av2m78	-3.00	1.07	-2.81	**0.0049**		
Soils Type 5	-268.51	128.80	-2.09	*0.0371*		
Soils Type 6	-155.62	114.92	-1.35	0.1757		
Soils Type 8	20.23	114.36	0.18	0.8596		
Soils Type 9	-119.87	118.95	-1.01	0.3136		
Soils Type 13	179.22	144.10	1.24	0.2136		
Soils Type 17	-76.64	119.51	-0.64	0.5214		
Soils Type 18	3.91	112.23	0.04	0.9722		
Soils Combo 32	435.74	120.13	3.63	**0.0003**		
Soils Combo 33	-509.33	157.27	-3.24	**0.0012**		
Soils Combo 35	153.67	119.94	1.28	0.2001		
Soils Combo 52	895.12	137.10	6.53	**0.0000**		
Soils Combo 53	379.34	132.10	2.87	**0.0041**		
Soils Combo 54	-157.06	115.49	-1.36	0.1739		

Table 42: Spatial error regression results and diagnostics for DstNclBCD Subset 2, model 3, spatial weights matrix 1

Variable	Coefficient	Std Error	z-value	Probability		
Intercept	6368.67	824.94	7.72	**0.0000**	Response Variable	DstNclBCD
Elevation	4.19	0.60	6.95	**0.0000**	Num of Observations	10,986
p_av2m34	-38.65	15.39	-2.51	*0.0120*	Pseudo-R^2	0.747
p_11	36.22	9.07	3.99	**0.0001**	Breusch-Pagan statistic	642.9
DSR_av2m78	-11.02	1.62	-6.83	**0.0000**	P-value	**0.0000**
Soils Type 5	-388.79	171.71	-2.26	*0.0236*		
Soils Type 6	-229.78	145.57	-1.58	0.1145		
Soils Type 8	7.11	149.34	0.05	0.9621		
Soils Type 9	-124.25	166.60	-0.75	0.4558		
Soils Type 13	122.85	191.24	0.64	0.5206		
Soils Type 17	50.26	159.37	0.32	0.7525		
Soils Type 18	179.03	142.37	1.26	0.2086		
Soils Combo 32	1075.51	170.65	6.30	**0.0000**		
Soils Combo 33	-257.50	185.53	-1.39	0.1652		
Soils Combo 35	294.79	155.15	1.90	0.0574		
Soils Combo 52	809.49	181.37	4.46	**0.0000**		
Soils Combo 53	505.50	180.72	2.80	**0.0052**		
Soils Combo 54	-194.68	152.95	-1.27	0.2031		
Lambda	0.87	0.01	150.84	**0.0000**		

Table 43: Spatial lag regression results and diagnostics for DstNclBCD Subset 3, model 1, spatial weights matrix 1

Variable	Coefficient	Std Error	z-value	Probability
W_DstNclBCD	0.84	0.01	138.86	**0.0000**
Intercept	271.87	306.97	0.89	0.3758
Elevation	1.88	0.24	7.66	**0.0000**
p_av3m345	-16.97	3.44	-4.94	**0.0000**
p_11	6.25	1.52	4.10	**0.0000**
t_8	20.88	15.99	1.31	0.1916
Soils Type 3	-30.98	116.78	-0.27	0.7908
Soils Type 6	-54.76	113.68	-0.48	0.6300
Soils Type 8	59.86	113.24	0.53	0.5971
Soils Type 9	-77.79	118.59	-0.66	0.5119
Soils Type 13	344.77	143.45	2.40	*0.0162*
Soils Type 17	-55.98	119.74	-0.47	0.6402
Soils Type 18	52.15	111.37	0.47	0.6396
Soils Combo 32	430.13	119.42	3.60	**0.0003**
Soils Combo 33	-474.23	157.50	-3.01	**0.0026**
Soils Combo 35	212.43	118.51	1.79	0.0731
Soils Combo 52	821.69	136.71	6.01	**0.0000**
Soils Combo 53	494.72	130.21	3.80	**0.0002**
Soils Combo 57	-278.01	118.76	-2.34	*0.0192*

Response Variable	DstNclBCD
Num of Observations	10,986
Pseudo-R^2	0.748
Breusch-Pagan statistic	497.2
P-value	**0.0000**

Table 44: Spatial error regression results and diagnostics for DstNclBCD Subset 3, model 1, spatial weights matrix 1

Variable	Coefficient	Std Error	z-value	Probability
Intercept	5043.17	1679.46	3.00	**0.0027**
Elevation	4.50	0.76	5.95	**0.0000**
p_av3m345	-79.31	16.81	-4.72	**0.0000**
p_11	51.80	8.47	6.11	**0.0000**
t_8	-170.98	96.89	-1.76	0.0776
Soils Type 3	11.90	156.06	0.08	0.9392
Soils Type 6	-129.50	139.09	-0.93	0.3518
Soils Type 8	-38.21	143.11	-0.27	0.7895
Soils Type 9	-98.78	162.16	-0.61	0.5424
Soils Type 13	321.51	186.74	1.72	0.0851
Soils Type 17	13.06	154.21	0.08	0.9325
Soils Type 18	135.97	135.39	1.00	0.3152
Soils Combo 32	817.40	165.18	4.95	**0.0000**
Soils Combo 33	-291.85	180.77	-1.61	0.1064
Soils Combo 35	259.03	147.81	1.75	0.0797
Soils Combo 52	750.67	175.82	4.27	**0.0000**
Soils Combo 53	669.66	173.24	3.87	**0.0001**
Soils Combo 57	-344.95	150.68	-2.29	*0.0221*
Lambda	0.87	0.01	151.40	**0.0000**

Response Variable	DstNclBCD
Num of Observations	10,986
Pseudo-R^2	0.751
Breusch-Pagan statistic	571.1
P-value	**0.0000**
Moran's I statistic	-0.0352

Table 45: Spatial lag regression results and diagnostics for DstNclBCD Subset 3, model 2, spatial weights matrix 1

Variable	Coefficient	Std Error	z-value	Probability
W_DstNclBCD	0.84	0.01	138.83	**0.0000**
Intercept	254.03	307.48	0.83	0.4087
Elevation	1.89	0.25	7.71	**0.0000**
p_av3m345	-16.85	3.44	-4.90	**0.0000**
p_11	6.26	1.52	4.10	**0.0000**
t_8	21.45	16.00	1.34	0.1801
Soils Type 3	-33.04	116.79	-0.28	0.7773
Soils Type 5	-231.52	127.67	-1.81	0.0698
Soils Type 6	-54.66	113.67	-0.48	0.6307
Soils Type 7	-338.37	133.44	-2.54	*0.0112*
Soils Type 8	60.56	113.24	0.53	0.5928
Soils Type 9	-76.90	118.59	-0.65	0.5167
Soils Type 13	340.58	143.50	2.37	*0.0176*
Soils Type 17	-57.46	119.75	-0.48	0.6313
Soils Type 18	53.51	111.37	0.48	0.6309
Soils Combo 32	432.59	119.44	3.62	**0.0003**
Soils Combo 33	-472.15	157.50	-3.00	**0.0027**
Soils Combo 35	213.26	118.51	1.80	0.0719
Soils Combo 52	815.15	136.87	5.96	**0.0000**
Soils Combo 53	497.00	130.23	3.82	**0.0001**

Response Variable	DstNclBCD
Num of Observations	10,986
Pseudo-R^2	0.748
Breusch-Pagan statistic	497.4
P-value	**0.0000**

Table 46: Spatial error regression results and diagnostics for DstNclBCD Subset 3, model 2, spatial weights matrix 1

Variable	Coefficient	Std Error	z-value	Probability
Intercept	5030.68	1679.47	3.00	**0.0027**
Elevation	4.51	0.76	5.97	**0.0000**
p_av3m345	-79.34	16.81	-4.72	**0.0000**
p_11	51.88	8.47	6.12	**0.0000**
t_8	-170.61	96.88	-1.76	0.0783
Soils Type 3	16.68	156.18	0.11	0.9150
Soils Type 5	-292.06	165.17	-1.77	0.0770
Soils Type 6	-130.05	139.09	-0.94	0.3498
Soils Type 7	-405.43	169.38	-2.39	*0.0167*
Soils Type 8	-37.99	143.10	-0.27	0.7907
Soils Type 9	-97.56	162.16	-0.60	0.5474
Soils Type 13	318.46	186.77	1.71	0.0882
Soils Type 17	11.78	154.22	0.08	0.9391
Soils Type 18	137.10	135.39	1.01	0.3112
Soils Combo 32	819.70	165.20	4.96	**0.0000**
Soils Combo 33	-288.99	180.80	-1.60	0.1100
Soils Combo 35	259.47	147.80	1.76	0.0792
Soils Combo 52	746.13	175.91	4.24	**0.0000**
Soils Combo 53	672.05	173.26	3.88	**0.0001**
Lambda	0.87	0.01	151.40	**0.0000**

Response Variable	DstNclBCD
Num of Observations	10,986
Pseudo-R^2	0.751
Breusch-Pagan statistic	571.5
P-value	**0.0000**
Moran's I statistic	-0.0353

Table 47: Spatial lag regression results and diagnostics for DstNclBCD Subset 3, model 3, spatial weights matrix 1

Variable	Coefficient	Std Error	z-value	Probability
W_DstNclBCD	0.84	0.01	138.78	**0.0000**
Intercept	300.81	307.32	0.98	0.3277
Elevation	1.92	0.25	7.82	**0.0000**
p_av3m345	-17.98	3.43	-5.25	**0.0000**
p_11	6.63	1.52	4.36	**0.0000**
t_8	20.38	16.00	1.27	0.2029
Soils Type 5	-224.51	127.73	-1.76	0.0788
Soils Type 6	-54.56	113.73	-0.48	0.6315
Soils Type 8	61.66	113.30	0.54	0.5863
Soils Type 9	-77.39	118.65	-0.65	0.5142
Soils Type 13	343.23	143.58	2.39	*0.0168*
Soils Type 17	-55.46	119.81	-0.46	0.6434
Soils Type 18	51.77	111.43	0.46	0.6422
Soils Combo 32	430.79	119.51	3.60	**0.0003**
Soils Combo 33	-477.13	157.58	-3.03	**0.0025**
Soils Combo 35	214.49	118.57	1.81	0.0705
Soils Combo 52	820.23	136.93	5.99	**0.0000**
Soils Combo 53	497.87	130.29	3.82	**0.0001**
Soils Combo 54	-114.82	114.11	-1.01	0.3143

Response Variable	DstNclBCD
Num of Observations	10,986
Pseudo-R^2	0.748
Breusch-Pagan statistic	495.4
P-value	**0.0000**

Table 48: Spatial error regression results and diagnostics for DstNclBCD Subset 3, model 3, spatial weights matrix 1

Variable	Coefficient	Std Error	z-value	Probability
Intercept	5123.11	1676.62	3.06	**0.0023**
Elevation	4.61	0.75	6.10	**0.0000**
p_av3m345	-81.01	16.78	-4.83	**0.0000**
p_11	52.14	8.46	6.16	**0.0000**
t_8	-172.26	96.72	-1.78	0.0749
Soils Type 5	-315.53	165.09	-1.91	0.0560
Soils Type 6	-128.94	139.17	-0.93	0.3542
Soils Type 8	-36.69	143.18	-0.26	0.7978
Soils Type 9	-91.96	162.24	-0.57	0.5708
Soils Type 13	326.77	186.86	1.75	0.0803
Soils Type 17	17.68	154.29	0.11	0.9088
Soils Type 18	133.08	135.46	0.98	0.3259
Soils Combo 32	818.42	165.28	4.95	**0.0000**
Soils Combo 33	-300.14	180.87	-1.66	0.0970
Soils Combo 35	257.51	147.89	1.74	0.0816
Soils Combo 52	749.30	176.00	4.26	**0.0000**
Soils Combo 53	667.74	173.35	3.85	**0.0001**
Soils Combo 54	-144.29	147.71	-0.98	0.3287
Lambda	0.87	0.01	151.07	**0.0000**

Response Variable	DstNclBCD
Num of Observations	10,986
Pseudo-R^2	0.750
Breusch-Pagan statistic	570.1
P-value	**0.0000**

Table 49: Spatial lag regression results and diagnostics for CSS Na2 Subset 2, model 1, spatial weights matrix 1

Variable	Coefficient	Std Error	z-value	Probability
W_CSSNa2	0.9593	0.0027	354.13	**0.0000**
Intercept	0.0692	0.0073	9.51	**0.0000**
Elevation	< -0.0001	0.0000	-2.83	**0.0047**
p_3	0.0001	0.0000	2.73	**0.0063**
p_9	-0.0002	0.0000	-4.06	**0.0001**
bio1	-0.0030	0.0006	-5.44	**0.0000**
Soils Type 3	0.0058	0.0025	2.38	*0.0175*
Soils Type 5	0.0065	0.0027	2.44	*0.0149*
Soils Type 6	0.0005	0.0024	0.21	0.8348
Soils Type 7	0.0057	0.0028	2.05	*0.0400*
Soils Type 8	-0.0040	0.0024	-1.69	0.0905
Soils Type 9	0.0019	0.0025	0.77	0.4442
Soils Type 13	-0.0018	0.0029	-0.62	0.5362
Soils Type 17	-0.0049	0.0025	-1.95	0.0510
Soils Type 18	-0.0026	0.0023	-1.13	0.2595
Soils Combo 32	-0.0040	0.0025	-1.60	0.1092
Soils Combo 33	0.0086	0.0033	2.61	**0.0090**
Soils Combo 35	0.0024	0.0025	0.95	0.3422
Soils Combo 52	-0.0046	0.0029	-1.60	0.1105
Soils Combo 53	0.0007	0.0028	0.25	0.8045

Response Variable	CSS Na2
Num of Observations	10,986
Pseudo-R^2	0.928
Breusch-Pagan statistic	337.3
P-value	**0.0000**

Table 50: Spatial error regression results and diagnostics for CSS Na2 Subset 2, model 1, spatial weights matrix 1

Variable	Coefficient	Std Error	z-value	Probability
Intercept	1.0881	0.0546	19.91	**0.0000**
Elevation	< -0.0001	0.0000	-1.46	0.1454
p_3	0.0013	0.0003	3.79	**0.0002**
p_9	-0.0028	0.0005	-5.40	**0.0000**
bio1	-0.0108	0.0050	-2.17	*0.0303*
Soils Type 3	0.0022	0.0034	0.64	0.5238
Soils Type 5	0.0055	0.0036	1.53	0.1265
Soils Type 6	0.0002	0.0031	0.06	0.9501
Soils Type 7	0.0039	0.0036	1.10	0.2732
Soils Type 8	-0.0054	0.0031	-1.74	0.0823
Soils Type 9	0.0029	0.0035	0.83	0.4066
Soils Type 13	-0.0013	0.0040	-0.33	0.7447
Soils Type 17	-0.0052	0.0034	-1.54	0.1225
Soils Type 18	-0.0050	0.0030	-1.69	0.0916
Soils Combo 32	-0.0106	0.0036	-2.93	**0.0034**
Soils Combo 33	0.0056	0.0039	1.43	0.1524
Soils Combo 35	0.0009	0.0033	0.27	0.7865
Soils Combo 52	-0.0058	0.0038	-1.51	0.1319
Soils Combo 53	0.0003	0.0038	0.07	0.9419
Lambda	0.9680	0.0025	392.29	**0.0000**

Response Variable	CSS Na2
Num of Observations	10,986
Pseudo-R^2	0.929
Breusch-Pagan statistic	329.2
P-value	**0.0000**

Table 51: Spatial lag regression results and diagnostics for CSS Nb2 Subset 2, model 1, spatial weights matrix 1

Variable	Coefficient	Std Error	z-value	Probability
W_CSSNb2	0.9591	0.0027	354.32	**0.0000**
Intercept	0.0720	0.0075	9.63	**0.0000**
Elevation	< -0.0001	0.0000	-3.43	**0.0006**
p_3	0.0001	0.0000	2.88	**0.0040**
p_9	-0.0002	0.0000	-4.06	**0.0001**
bio1	-0.0033	0.0006	-5.71	**0.0000**
Soils Type 3	0.0059	0.0025	2.33	*0.0200*
Soils Type 5	0.0064	0.0028	2.31	*0.0210*
Soils Type 6	0.0003	0.0025	0.13	0.8967
Soils Type 7	0.0059	0.0029	2.04	*0.0416*
Soils Type 8	-0.0044	0.0025	-1.78	0.0744
Soils Type 9	0.0019	0.0026	0.75	0.4520
Soils Type 13	-0.0032	0.0030	-1.06	0.2876
Soils Type 17	-0.0057	0.0026	-2.18	*0.0290*
Soils Type 18	-0.0031	0.0024	-1.30	0.1934
Soils Combo 32	-0.0052	0.0026	-2.03	*0.0425*
Soils Combo 33	0.0077	0.0034	2.28	*0.0229*
Soils Combo 35	0.0015	0.0026	0.59	0.5539
Soils Combo 52	-0.0047	0.0029	-1.60	0.1089
Soils Combo 53	0.0002	0.0029	0.08	0.9324

Response Variable	CSS Nb2
Num of Observations	10,986
Pseudo-R^2	0.928
Breusch-Pagan statistic	326.9
P-value	**0.0000**

Table 52: Spatial error regression results and diagnostics for CSS Nb2 Subset 2, model 1, spatial weights matrix 1

Variable	Coefficient	Std Error	z-value	Probability
Intercept	1.0759	0.0566	19.01	**0.0000**
Elevation	-0.0001	0.0000	-1.76	0.0790
p_3	0.0013	0.0003	3.73	**0.0002**
p_9	-0.0026	0.0005	-4.99	**0.0000**
bio1	-0.0108	0.0051	-2.10	*0.0359*
Soils Type 3	0.0019	0.0035	0.55	0.5840
Soils Type 5	0.0045	0.0037	1.21	0.2259
Soils Type 6	-0.0006	0.0032	-0.20	0.8377
Soils Type 7	0.0030	0.0037	0.82	0.4112
Soils Type 8	-0.0062	0.0032	-1.92	0.0544
Soils Type 9	0.0027	0.0036	0.74	0.4570
Soils Type 13	-0.0031	0.0041	-0.75	0.4550
Soils Type 17	-0.0068	0.0035	-1.96	*0.0500*
Soils Type 18	-0.0063	0.0031	-2.03	*0.0421*
Soils Combo 32	-0.0134	0.0037	-3.58	**0.0004**
Soils Combo 33	0.0033	0.0040	0.83	0.4057
Soils Combo 35	-0.0013	0.0034	-0.39	0.6963
Soils Combo 52	-0.0058	0.0039	-1.46	0.1445
Soils Combo 53	-0.0009	0.0039	-0.22	0.8290
Lambda	0.9683	0.0025	394.74	**0.0000**

Response Variable	CSS Nb2
Num of Observations	10,986
Pseudo-R^2	0.928
Breusch-Pagan statistic	319.4
P-value	**0.0000**






Historic England Research and the Historic Environment

We are the public body that looks after England's historic environment. We champion historic places, helping people understand, value and care for them.

A good understanding of the historic environment is fundamental to ensuring people appreciate and enjoy their heritage and provides the essential first step towards its effective protection.

Historic England works to improve care, understanding and public enjoyment of the historic environment. We undertake and sponsor authoritative research. We develop new approaches to interpreting and protecting heritage and provide high quality expert advice and training.

We make the results of our work available through the Historic England Research Report Series, and through journal publications and monographs. Our online magazine Historic England Research which appears twice a year, aims to keep our partners within and outside Historic England up-to-date with our projects and activities.

A full list of Research Reports, with abstracts and information on how to obtain copies, may be found on www.HistoricEngland.org.uk/researchreports

Some of these reports are interim reports, making the results of specialist investigations available in advance of full publication. They are not usually subject to external refereeing, and their conclusions may sometimes have to be modified in the light of information not available at the time of the investigation.

Where no final project report is available, you should consult the author before citing these reports in any publication. Opinions expressed in these reports are those of the author(s) and are not necessarily those of Historic England.

The Research Report Series incorporates reports by the expert teams within the Research Group of Historic England, alongside contributions from other parts of the organisation. It replaces the former Centre for Archaeology Reports Series, the Archaeological Investigation Report Series, the Architectural Investigation Report Series, and the Research Department Report Series

ISSN 2398-3841 (Print)
ISSN 2059-4453 (Online)