THE ATLAS OF RURAL SETTLEMENT IN ENGLAND GIS

Documentation

Andrew G Lowerre

If you would like this document in a different format,
please contact our Customer Services department:
Telephone: 0870 333 1181
Textphone: 01793 414878
E-mail: customers@english-heritage.org.uk

INTRODUCTION

This document contains information about the Atlas of Rural Settlement in England GIS project, the nature and limitations of the data in the data collection produced, and the processes by which the data and accompanying metadata were created. Users of the data collection are strongly encouraged to read the information carefully in order to understand the origins, contents, strengths and weaknesses of the data.

Aims and objectives of the conversion project

The aim of the project was to enable the key maps of rural settlement (figs 13, 15 and 17) and terrain (figs 14, 16 and 18) presented in Brian K Roberts and Stuart Wrathmell's *An Atlas of Rural Settlement in England* (2000) to be used more effectively in future research on landscape and settlement in England, as well as in the management of the historic environment. The maps printed in the *Atlas* were produced digitally, but were created as vector graphics files, and were therefore not useable in Geographic Information Systems (GIS) software. Given the now-widespread use of GIS software in the management and study of the historic environment, as well as the availability of software such as Google Earth, that lacuna significantly restricted the use and value of the *Atlas*'s maps.

The project had three objectives supporting the overall aim:

- to convert the original graphics files from which the published maps were printed into geo-referenced spatial and attribute data and to create accompanying metadata and documentation for the resulting datasets;
- to archive the datasets, metadata and documentation and disseminate them to all interested parties via English Heritage's National Monuments Record; and
- to publicise the creation of the datasets and their availability to interested parties through a carefully targeted series of publications and presentations at one or more conferences.

It is hoped that the creation and dissemination of the data described here will stimulate future work relating to the topics discussed and questions posed in the printed *Atlas*. Presenting Roberts and Wrathmell's materials in an interactive, spatially-aware digital format will enable a variety of users to examine, query and re-interpret Roberts and Wrathmell's results. The research potential of combining the *Atlas* data with a wide range of other regional and national datasets is enormous. Some exploratory analyses and re-visualisations of the data described here are presented in a forthcoming article in the journal *Landscapes* (Lowerre forthcoming).

People involved in the project

The project was carried out with the agreement and enthusiastic support of Prof Roberts and Dr Wrathmell. The project also benefited from the advice and encouragement of English Heritage's Characterisation Team, in particular Graham Fairclough, David Stocker and Roger M Thomas.

Eddie Lyons converted the graphics files supplied by Brian Roberts into a format readable in ArcGIS. Kirsty Stonell Walker scanned the pages of the printed *Atlas* containing the settlement province and sub-province descriptions. Sheila Keyte processed the scanned

pages of the *Atlas* using Optical Character Recognition (OCR) software to create a digital version of the texts. All other tasks were carried out by Andrew Lowerre.

Viewing the data

The data collection is intended to be largely platform-independent. The spatial and attribute data are presented in two different formats: ESRI shapefile and Google/Open Geospatial Consortium KMZ. The ESRI shapefile format can be read by most leading proprietary (eg ArcGIS and MapInfo) and open-source (eg GRASS and Quantum GIS) GIS software packages. ESRI shapefiles can also be read by recent versions of many leading CAD (Computer-Aided Drafting/Design) software packages. The shapefiles can also be viewed in free GIS 'data viewers' such as ESRI's ArcGIS Explorer. The KMZ format – a compressed version of Keyhole Markup Language or KML – is most often viewed in 'geobrowser' software such as Google Earth, and can also be accessed using ESRI's ArcGIS Explorer. Those wishing to make use of the Atlas of Rural Settlement in England GIS data collection who do not already have access to GIS, CAD or geobrowser software are advised to search the Internet for software suited to their needs.

It should be noted that mention of a specific software package in this or any other documentation or metadata describing the Atlas of Rural Settlement in England GIS data collection does not constitute or imply an endorsement by English Heritage of that software package or vendor.

Preferred citation

All works which use or refer to the materials in the Atlas of Rural Settlement in England GIS data collection should acknowledge the data collection as a source by means of bibliographic citation. The preferred bibliographic citation for this data collection is:

Lowerre, A G, Lyons, E R, Roberts, B K, and Wrathmell, S 2011 *The Atlas of Rural Settlement in England GIS: Data, Metadata and Documentation* [computer file]. Swindon: English Heritage

CONTENTS OF THE DATA COLLECTION

The data collection comprises the following elements:

- spatial and attribute data, supplied in two different formats: ESRI Shapefile and Google/Open Geospatial Consortium KMZ;
- an Adobe Portable Document Format (PDF) file called 'AtlasRuralSettlementEnglandGIS_ProvincialSubProvincialDescriptions.pdf', which contains text and figures for the settlement province and sub-province descriptions found on pages 40–57 of the printed *Atlas*;
- ArcGIS 'layer' definition files (*.lyr) – compatible with ArcGIS versions 9.0 and above – recording suggested symbolisation for the various shapefiles in the collection; these symbolisations were used when creating the KMZ files;
- an ArcGIS map document (*.mxd) – compatible with ArcGIS versions 9.0 and above – depicting the shapefiles in the data collection, using the symbolisation recorded in the *.lyr files noted above;

- UK GEMINI version 2.1-compliant discovery level metadata in XML version 1.0 format for the spatial and attribute data;
- an Adobe Portable Document Format (PDF) file called 'AtlasRuralSettlementEnglandGIS_DataDictionary.pdf', a data dictionary detailing the attribute field names, suggested aliases and descriptions of the types of data held in each field; and
- an Adobe Portable Document Format (PDF) file called 'AtlasRuralSettlementEnglandGIS_Documentation.pdf' (the current document).

NATURE AND LIMITATIONS OF THE DATA

It is essential that users of the data collection understand the methods by which the original maps were created, how those maps were transformed into spatial and attribute data usable in GIS and similar software, and the limitations of the data arising from the manner of their creation.

The nature of the data

Throughout the documentation and metadata included in the Atlas of Rural Settlement in England GIS dissemination package, the word 'data' is used to refer to the shapefiles and KMZ files that store the spatial representations and accompanying attributes of features depicted on maps printed in Roberts and Wrathmell's published *Atlas*. It must be emphasised strongly, however, that the data are 'data of interpretation' and not unmediated, primary or purely empirical 'facts' in themselves.

Roberts and Wrathmell describe the process by which they created the printed maps on pages 9–17 of the *Atlas*. The process was one of interpretation and characterisation of the landscape of England at a national scale, using the nineteenth-century Ordnance Survey 'Old Series' 1:63,360 (one inch to one mile) scale maps as a source. The delineation of settlement provinces, sub-provinces and local regions was based on a carefully-reasoned but nonetheless subjective method, involving, as Roberts and Wrathmell put it, 'little science but much logic' (Roberts and Wrathmell 2000, 13). Similarly, the maps of terrain are a highly generalised, synthetic characterisation of the physical landscape of England, based on a multitude of sources. Roberts and Wrathmell state that the maps, explanations and analyses presented in the printed *Atlas* should be understood 'not as definitive statements on regional diversity, but rather, as an initial attempt to provide an alternative perspective on historic regional variation' (ibid, vii). The spatial and attribute data in this collection should be understood and used in the same vein.

Questions of scale

The locations of points and polygon and line vertices in the data are stored at sub-metre *precision*, meaning that the X- and Y-coordinates are recorded to the nearest millimetre. The processes by which the original graphics files were produced, however, mean that the positional *accuracy* of the source maps and thus the data derived from them is 1,000 metres at best. Roberts and Wrathmell based their work on the Old Series one-inch maps, but transcribed the nucleation and dispersion information first onto 1:250,000 road atlases, and then onto a 1:1 million base map in their graphics software. The Atlas GIS data are best suited to giving a national or regional picture or to putting a local- or

county-scale study into a wider perspective. Displaying the Atlas data at the scale of, say, a single parish will likely give unsatisfactory results. GIS and similar software make it possible to display the data at any scale, but users zoom in beyond a scale of about 1:200,000 at their own risk.

Coastline and national borders

The coastline of England and Wales and the national borders between England, Wales and Scotland used in the maps in the printed *Atlas* were adapted from a 1928 Ordnance Survey map. These representations of the coast and the borders have been kept in the GIS data. Users of the data collection will see that the coastline and national boundaries do not line up perfectly with more modern, and arguably more precise and accurate spatial data. This is to be expected, given the processes by which Roberts and Wrathmell created their original maps. The coastline and the Welsh and Scottish borders could have been massaged to fit better with more modern data (eg, smaller-scale mapping and data now freely available via the Ordnance Survey's 'OpenData' programme (Ordnance Survey 2010)), but a conscious decision was taken not to do so. To 'fix' some elements of the data would lend a spurious air of precision to the rest.

Also, it is anticipated that different users of the data collection will deploy the data to different purposes. For a researcher working on nineteenth-century rural settlement, it would be preferable to match the Atlas data to mid-nineteenth-century versions of the national borders. But for a user most interested in viewing Roberts and Wrathmell's results in the context of understanding rural settlement in the early twenty-first century, 'fixing' the Welsh and Scottish borders to match the current (2010) ones would be more appropriate. Users who wish to modify the data to best suit their needs are encouraged to do so.

Locations of Nucleation points

Like the coastline, the points in the data representing nucleated settlements – the towns, villages and hamlets that dot the countryside – will not match perfectly with more modern, larger scale data. The positions of the nucleation points relative to each other are accurate enough when working at a national scale, but users will find, as they zoom in, that the locations of nucleation points may appear out of place relative to other data. Again, this is because of the processes by which the original maps were produced. It was beyond the scope of the conversion project to check the location of every nucleation point against modern or historic Ordnance Survey mapping.

Solidity and location of boundaries

The precision in the data of the boundaries between one sub-province and the next or between two terrain zones should not be misinterpreted. Because of the way GIS data are stored, there are clearly delineated edges between polygons, but these can give the erroneous impression that there are sudden changes in landform or settlement patterns from one area to another, particularly when viewed at larger scales. Such changes can, of course, be quite subtle and dispersed over a considerable geographic extent. It is worth quoting Roberts and Wrathmell regarding the solidity and location of the boundaries they mapped: '[i]t should be appreciated that in all of our maps the drawn boundary forms a band approximately one and a half to two kilometres in width: while the observant traveller would detect the landscape changes when crossing this zone, on-ground

definition of a line can be difficult if not wholly impossible, and any such boundary may resolve itself into a narrow and complex transitional zone rather than a thin line' (ibid, 45). The same caveat applies to the Atlas GIS data.

FROM GRAPHICS FILES TO SPATIAL AND ATTRIBUTE DATA

This section describes in detail the processes by which the graphics files used in the production of maps in the printed *Atlas* were converted to shapefile- and KMZ-format spatial and attribute data. The conversion of the graphics files to AutoCAD drawing (*.DWG) format and initial cleaning and georeferencing of the AutoCAD drawings was undertaken by Eddie Lyons. The conversion of the data from AutoCAD drawing format to shapefile and KMZ format was undertaken by Andrew Lowerre. The work was carried out using Adobe Illustrator and PhotoShop CS2, Autodesk AutoCAD Civil 3D 2007 and 2008 and ESRI ArcGIS 9.1, 9.2 and 9.3. In ArcGIS, functions included in the ET GeoTools 9.4 (Tchoukanski 2008a) and ET GeoWizards 9.8 (Tchoukanski 2008b) extensions were used in addition to the standard, 'off the shelf' tools.

The conversion processes

The methods used to convert the different elements in the data collection from the graphics files to GIS-ready data varied. The descriptions below take each element or group of elements in turn.

Polygon data

The process of converting the original FreeHand files to GIS polygon data for the Terrain Types and Zones and settlement Provinces, Sub-provinces and Local Regions/Dispersion Zones was essentially the same. The FreeHand format files were converted by Brian Roberts into Adobe Illustrator (*.AI) format files. The *.AI files were opened in Adobe Illustrator CS2 and exported as AutoCAD drawing files.

The *.DWG files were opened in AutoCAD and unwanted objects (eg, hatch pattern fills, county boundaries and unneeded cartographic furniture) were deleted. Spline objects representing the outlines of the terrain zones, the coastline and the internal borders with Wales and Scotland were retained. To convert these to polylines (the version of ArcGIS used does not recognise AutoCAD spline objects) the following procedure was used:

1. Each drawing file was saved to an AutoCAD R12 .DXF format (this converts splines to 3D polylines);
2. AutoCAD Civil 3D's Drawing Cleanup tool was used to simplify the 3D polylines (reducing large numbers of nodes that approximated the shapes of the splines) and convert them to 2D polylines (using Simplify Objects, with a tolerance value of 5);
3. Each file was saved again as an AutoCAD drawing file.

The drawing files were georeferenced individually in AutoCAD to the British National Grid, based on the grid lines drawn originally in the FreeHand illustrations. These georeferenced drawing files were then combined to create a single drawing file. A degree of overlap existed between the individual files. In the areas of overlap between the components duplicate linework was edited out, with linework from the south-east component retained in preference.

The polyline features from the AutoCAD drawing files were loaded into an ArcGIS map document and exported to new polyline feature classes in an ArcGIS 9.2 format personal geodatabase using the ArcGIS 'Export' tool.

In ArcGIS, it was necessary to clean the polyline feature classes to prepare them to be used to build the polygon features representing the terrain and settlement province, sub-province and local region/dispersion zone data. Standard ArcGIS and ET GeoTools editing tools were used to remove a great deal of redundant linework, pseudo nodes and dangling polylines in the data. The 'dirty' data were present because the linework in the original graphics files had not been digitised to create complete, closed polygons defining each area or to 'snap' the vertices of one outline to the vertices of its neighbours. And while the hatch *patterns* in the original graphics files had been removed in AutoCAD, the *outlines* of the graphic hatches came through into the ArcGIS data as additional, individual lines. The outlines of the hatches almost invariably did not match precisely the outlines of the terrain or dispersion zone areas, and so the different sets of linework had to be painstakingly unpicked.

Topology for the linework was defined to enable the creation of clean polygons from the polyline feature classes. A cluster tolerance of 0.001metres was used and the topology ensured that the linework did not have any dangles, did not have pseudo nodes, did not self-overlap, did not self-intersect, that all lines were single-part and that no lines overlapped. To ensure that the coastline and the borders with Wales and Scotland would match across all of the final polygon data, the linework depicting the coast and the borders was copied from the local region/dispersion zone polyline feature class and applied to the terrain data as well.

Once all the linework was topologically clean, the ArcGIS tool 'Feature to Polygon' was used to build new polygon feature classes for the terrain zones and the settlement local regions/dispersion zones. The approach used was to create the smallest, most detailed polygons first – the terrain zones and settlement local regions/dispersion zones – and then aggregate the more detailed polygons together to build up the larger, overarching areas – the terrain types and settlement sub-provinces and provinces. The polygon feature classes were checked, and where necessary edited, to confirm that there were no overlapping polygons and no inappropriate gaps between them. The only exceptions to the general rule that the polygons must not overlap were the features representing terminal moraines and drumlins in the terrain zones feature class. The moraines and drumlins do overlie other terrain zone polygons.

Fields were added to the Terrain Zone polygon feature class to hold the attribute values for the name of the terrain zone (eg, Carboniferous Limestone landscapes or outwash sands and gravels) as well as the terrain type (ie, Uplands, Intermediate Lands or Lowlands). Raster images of the terrain maps were created from the Adobe Illustrator files, and these were loaded into an ArcGIS map document, georeferenced and used as a backdrop to assign the appropriate attribute values to each Terrain Zone polygon. In a few instances, the terrain zone could not be determined from the original maps, and values were assigned with the help of Brian Roberts.

Fields were added to the local region/dispersion zone polygon feature class to record the attribute values for density of dispersion, the name and code of the local region, the name and code of the sub-province, and the name of the province. Attribute values for the relevant province and sub-province were assigned to each polygon using Figure 1 in the printed *Atlas* as a reference. The local region names and codes listed on pages 67–9 in the printed *Atlas* was used as a basis for assigning local region values to the various polygons. The printed *Atlas* does not, however, include a map indicating which local region name and code applies to which area. Brian Roberts supplied such a map, making it possible to assign local region values to most of the polygons. Where inconsistencies or uncertainties still remained, slight changes were made in consultation with Brian Roberts to a few polygons' geometry, as well as to the names and alphanumeric codes applied to a small number of local settlement regions. As a result, the boundaries of the local regions and the list of local region names and codes in the final GIS data do not match exactly those presented in the printed *Atlas*. The differences between the GIS data and the printed *Atlas* are, however, neither substantive nor meaningful.

To enable the capture of density of dispersion information for each polygon, Eddie Lyons exported a version of Figure 3 from the *Atlas* from Adobe Illustrator format to a raster image. In Adobe PhotoShop CS2, a new colour ramp was applied to the areas of dispersion in order to highlight more clearly the differences between the fourteen colour classes. This was done to ease the visual differentiation of one dispersion zone from another. The re-coloured raster image of Figure 3 from the *Atlas* was loaded into an ArcGIS map document and georeferenced. The image was used as a background against which to capture the dispersion attributes for each polygon.

In many cases, local regions are characterised with a single description of the degree of settlement dispersion: the Cheviot Margin (CWRTD7) has very low densities of dispersion, and Macclesfield Forest (WCHPL4) has high to very high densities. There are, however, instances where a Local Region may be comprised of two or more polygons with different dispersion values. The Lower Thames (ETHAM1) is comprised of polygons with both extremely low to very low densities and very low to low densities, and the Carlisle Coast (WCUSL2) has both very low to low and low to medium densities of dispersion.

Once all the attribute values for the lowest-level polygons – the terrain zones and settlement local regions/dispersion zones – had been applied, it was possible to build the higher-level polygons from them. The ArcGIS 'Dissolve' tool was used to combine and merge the local region/dispersion zone polygons based on their sub-province values, creating a new polygon feature class depicting the sub-provinces. The same method was used to construct the settlement province polygon feature class and to create the terrain types feature class from the terrain zones polygons. The polygon feature class for the 'background' layer depicting England and Wales was created by dissolving the settlement provinces.

Terrain Scarps

In the AutoCAD terrain drawing, linework representing scarps was selected and placed on a new layer named 'Scarps'. Extra linework that was used in FreeHand to act as masks or as stylised conventions, and which were extraneous to the data requirements, were deleted.

In ArcGIS, all features in the AutoCAD drawing on the 'Scarps' layer were exported to a new polyline feature class in an ArcGIS 9.2 format personal geodatabase using the ArcGIS 'Export' tool. This feature class was symbolised to imitate the depiction of scarps in the printed maps, ie with 'dangling hatches'. All scarp lines were examined to ensure that the direction of the hatches on the lines matched that in the printed maps. Where necessary, the ET GeoTools 'Flip' tool was used to change the vertex order of the lines so the 'dangling hatches' would dangle in the correct direction. The topology of the scarps polyline feature class was checked to ensure that lines did not have pseudo nodes, did not self-overlap, did not self-intersect, were single-part and did not overlap with each other.

Nucleations

Three FreeHand format files (corresponding to figs 13, 15 and 17 in the printed *Atlas*) were converted by Brian Roberts into Adobe Illustrator (*.AI) format files. The *.AI files were opened in Adobe Illustrator CS2 and exported as AutoCAD drawing files.

The three *.DWG files were opened in AutoCAD and unwanted objects (eg, hatch pattern fills, dispersion zone outlines, county boundaries and unneeded cartographic furniture) were deleted. The circular symbols representing the nucleations and spline objects representing the coastline and the internal borders with Wales and Scotland were retained.

The three drawing files were georeferenced individually in AutoCAD to the British National Grid, based on the grid lines drawn originally in the FreeHand illustrations. These three georeferenced drawing files were then combined to create a single drawing file. A degree of overlap existed between the three individual files, with obvious duplication of the nucleation symbols. The total number of symbols (AutoCAD circle objects) at this stage was 11,738. The gridlines were deleted after georeferencing.

The overlaps revealed slight differences in the georeferencing between the three individual AutoCAD drawing files, based on the limited accuracy of the grid lines drawn in the FreeHand files. Because the south-east component file included the greatest number of symbols (6,676) these were retained in preference. Duplicate symbols in the overlap regions from the north component and the south-west component were deleted. These duplicates were identified visually, and selected and deleted manually.

AutoCAD also showed that the five sizes of symbol used for the original maps did not export as identical sizes of circle; each one varied slightly in diameter. AutoCAD's object selection filters were used to select globally each of the five ranges of symbol sizes and the object Properties pallet was used to apply a uniform diameter to each range. Five symbol sizes were applied (from largest to smallest, diameter in AutoCAD units, which equate to metres on the British National Grid): 3000, 2500, 2000, 1500, 1000.

Each range of symbols were placed on new AutoCAD layers, respectively from largest to smallest: A, B, C, D and E. These equate to the categories given on page 11 in the printed *Atlas*. After the duplications in the overlap areas were deleted, the total number of symbols remaining was 10,956. A total of 10,963 is given in the printed *Atlas* (Roberts and Wrathmell 2000, 11). AutoCAD Civil 3D's 'Drawing Cleanup' tool also showed that there were 418 duplicates in the dataset.

The AutoCAD drawing with the combined nucleation symbols was then loaded into an ArcGIS map document. The circle symbols were interpreted in ArcGIS as polygon features. All features were exported to a new polygon feature class in an ArcGIS 9.2 format personal geodatabase using the ArcGIS 'Export Data' tool. A new point feature class was created from the polygon circles using the ArcGIS tool 'Feature to Point' set to create centroid points inside each polygon.

The ET GeoWizards 9.8 tool 'Remove Exact Duplicates' was used to remove the duplicate points noted in the AutoCAD drawing. It was then noted that, after having removed the exact duplicates, there were a few visible instances in the point feature class where there were points extremely close together, too near to represent separate nucleations. Comparison with georeferenced raster images of the nucleation maps confirmed that there were cases where two points existed where only one nucleation was represented on the published map. A small number of these duplicate points were located through visual inspection and deleted. As a further check, the ArcGIS tool 'Near (Analysis)' was run to find the distance from each point to the next nearest point in the same feature class, using a search radius of 1,000m. The overwhelming majority of points did not have a neighbour within a radius of 1,000m. Forty-four points had nearest neighbours less than 1m away. These were clustered in an area where the south-western and south-eastern nucleations layers from the three initial AutoCAD drawings overlapped. These were simply a small number of duplicate symbols that had not been removed in the AutoCAD drawing. The first of each pair of 'near duplicate' points was deleted by hand. The number of nucleation points in each category stated by Roberts and Wrathmell in the printed *Atlas*, the number in the final GIS data, and the difference between the two sums are presented in the following table:

	Number of Nucleations		
Nucleation Category	Stated in printed *Atlas*	In final point data	Difference
A	263	249	14
B	635	600	35
C	2782	2684	98
D	4841	4710	131
E	2442	2270	172
Total	10,963	10,513	450

Legacy fields derived from the original import of data from the AutoCAD drawing file (eg, 'Entity', 'Handle', 'Color', 'Linetype', etc.) were deleted. The field 'Layer' was initially retained, as this was where the nucleations categories from the Atlas were held in the AutoCAD drawing file. A new text field, 'NuclCat_A' (for Nucleation Category – Alphabetic), was created and the values from 'Layer' were copied into it. The 'Layer' field was then deleted.

A new numeric (short integer) field, 'NuclCat_N' (for Nucleation Category – Numeric), was also added. Integer values ranging from 5 to 1 were applied to each point feature, the integer value corresponding to the nucleation category letter derived from the Atlas, as shown in the following table:

Nucleation Category – Alphabetic	Nucleation Category – Numeric
A	5
B	4
C	3
D	2
E	1

The integer values were added in order to enable easy representation of the data using graduated symbols in GIS software. It must be emphasised that the numeric values for the nucleation categories are *rank* data, derived from a subjective categorisation by Roberts and Wrathmell of their source. These values should not be used to perform calculations – two category D/2 nucleations do not 'equal' one category B/4 nucleation.

The locations of the nucleation points were checked to see that they all lay within the polygons depicting the settlement provinces. Where necessary, points were moved just enough to ensure this was the case. Comparison with figure 17 in the printed *Atlas* indicated that eight nucleations in the original maps were depicted lying west of the English–Welsh border (ie, outside the settlement province polygons). These nucleations were not moved.

Dispersion Scores/Hamlet Counts

One Adobe Illustrator (*.AI) format file was provided by Brian Roberts. The content of the file was similar to figure 7 in the printed *Atlas*, but covered the whole of England. The file was opened in Adobe Illustrator CS2 and exported as an AutoCAD drawing file (*.DWG). The annotation and polygon layers from the AutoCAD drawing were loaded into an ArcGIS map document. These layers represented, respectively, the positioned text of the dispersion scores and hamlet counts and the circles highlighting those scores/counts which Roberts and Wrathmell noted as 'unusual'.

A point feature class was created from the centroids of the 'highlight' circles using the ArcGIS 'Feature to Point' tool. Another point feature class was created from the AutoCAD drawing using the ESRI sample tool 'CADtoFeatureClass', specifying that only the annotation class from the drawing be converted to a point feature class. This created a series of pairs of points, one representing the dispersion score at each sample location, the other the hamlet count.

The ET GeoWizards 9.8 tool 'Remove Exact Duplicates' was used to remove exact duplicate points from the dispersion score/hamlet count points. To check for 'near duplicates' (ie, points that were extremely close together but not coincident), the ArcGIS 'Near' tool was run, using a search radius 2,500m. Two sets of pairs of points were identified where the features were very close (ca. 2.5 to 3.5m distant). In each case, the text values of the very near points were the same, so one of each 'very near' pair was deleted.

One point (feature ID 3878 in the final GIS data) which records a hamlet count but did not have a corresponding dispersion score was identified. Two points (feature IDs 3879

and 3880 in the final GIS data) which appeared to record dispersion scores which did not have corresponding hamlet counts were also identified. The dispersion scores and hamlet counts were stored as separate points, with the score/count attribute stored in a 'Text_' field. Examination of the values in this field highlighted six records with apparently erroneous values. These values were changed to what seemed the most likely correct value. The erroneous values, the number of records with those values, and the changes applied are listed in the following table:

Erroneous Value	Number of Records	Changed To
3H3	1	H3
H00-3	1	0-3
H05	1	5
HO	2	H0 [zero]
Q	1	1

Changing 'Q' to '1' was done on the assumption that the 'Q' was a typographical error for either 1 or 2. Using the principle 'if in doubt round down' applied during the creation of the original maps (see Roberts and Wrathmell 2000, 12), the lower value 1 was chosen.

All the hamlet count values (ie, those values in 'Text_' that contained an 'H') were transferred into a new field, called 'Hamlet_Count'. Separate point feature classes for the dispersion scores and the hamlet counts were then created by selecting and exporting first, all those features with dispersion scores (ie, all those values remaining in the field 'Text_'), then all those features with hamlet counts.

Comparison of the locations of the points in the dispersion score and hamlet count feature classes on the one hand and the centroid points of the highlight polygons on the other indicated that the points in the dispersion score and hamlet count feature classes were offset to the north and west of the centroid points. This was the result of how ArcGIS placed the original dispersion score and hamlet count points from the AutoCAD annotation when the data were initially imported. It was assumed that the centroid points best represented the sample locations originally mapped by Roberts and Wrathmell. It was necessary to concatenate the dispersion score and hamlet count points so that a single point – carrying both the dispersion score and hamlet count attributes – would represent Roberts and Wrathmell's sample locations, and then reposition the merged points to match the centroid points.

The first step was to move all the dispersion score points 900m to the south, using the ET GeoWizards 9.8 tool 'Move Shapes'. This was done to ensure that, in every instance, the point representing a dispersion score was closest to the point representing its corresponding hamlet count. Otherwise, it would have been possible, for example, for the hamlet count point nearest a given dispersion score point to have belonged to a different, adjacent pair of points. The ArcGIS tool 'Integrate' was used to make coincident those points in the dispersion score and hamlet count feature classes that lay within a distance of 300m of each other. This process was checked with the ArcGIS 'Near' tool, using a search radius of 1,000m, which showed that the distance from each point in the

dispersion score feature class to the nearest point in hamlet count feature class was in all cases zero metres.

The two separate point feature classes were then combined, using the ArcGIS tool 'Intersect', producing a new point feature class, which recorded both the dispersion score and the corresponding hamlet count for each individual point. The two dispersion score points with no corresponding hamlet count points and the one hamlet count point with no corresponding dispersion score were added to the new, combined point feature class by hand. New text fields called 'Disp_Scr_A' (for 'Dispersion Score' (alphabetic)) and 'Ham_Cnt_A' (for 'Hamlet Count' (alphabetic)) were created and the appropriate attribute values copied into them from the existing attributes. Extraneous fields were then deleted.

In order to move the points in the integrated dispersion score/hamlet count feature class to locations more closely coinciding with the sample locations represented in the highlight polygon centroid point feature class, it was necessary to determine the extent to which the XY coordinates of each 'unusual' point in the integrated dispersion score/hamlet count feature class differed from those of the centroid points. The X and Y coordinates of each point in the integrated dispersion score/hamlet count feature class that lay within a highlight polygon were compared to the X and Y coordinates of the centroid of each highlight polygon. The average differences in X and Y coordinates (649.6712 and 25.06032, respectively) were calculated.

New X and Y coordinates for each point were generated by adding 649.6712 to the current X value and 25.06032 to the current Y value, then rounding the values to the nearest 100m, and adding 55m to the rounded number. This gave each point coordinates for the centre of a 100x100m square nearest the coordinates supplied by adding the average X and Y differences. The end result was to move all the points such that those that were highlighted with 'unusual' circles in the original graphics file were located very close to the centre of each circle, while moving all the rest of the points by the same amount. The locations of the dispersion score/hamlet count points were checked to see that they all lay within the polygons depicting the settlement provinces. Where necessary, points were moved just enough to ensure this was the case.

A new text field, 'Unusual_A', was created to indicate those points which were circled as representing unusual values in the original graphics file. Using the spatial selection tools in ArcGIS, those points in the final dispersion score/hamlet count feature class which lay completely within the 'highlight' circles were selected. The values in field 'Unusual_A' for those points were set to 'Y'; the value for all other points was set to 'N'.

The dispersion score and hamlet count values stored in fields 'Disp_Scr_A' and 'Ham_Cnt_A' are stored as text data, rather than numeric data. In order to enable the use of many spatial analysis and interpolation tools provided by GIS software on the data, it was necessary to store these data as numeric values. In most instances, the dispersion score values stored as text in 'Disp_Scr_A' were simply numbers, eg, 1, 2, 3, 5, 8 and so on, as outlined in the *Atlas* (Roberts and Wrathmell 2000, 12) . In 260 cases, however, the values in the field 'Disp_Scr_A' included ranges such as '0–1', '1–3', and '3–8'. Three new numeric field were created to hold different permutations of these 'range' values: 'Disp_Scr_N1', 'Disp_Scr_N2' and 'Disp_Scr_N3'. These fields were used to store,

respectively, the high, medium and low values in the ranges held in 'Disp_Scr_A'. The 'range' values encountered in the data and the numeric values entered in the three fields are shown in the following table:

Original 'Range' Value	Instances	Disp_Scr_N1 (High)	Disp_Scr_N2 (Medium)	Disp_Scr_N3 (Low)
0–1	197	1	0	0
0–2	37	2	1	0
0–3	4	3	1	0
1–0	2	1	0	0
1–2	10	2	1	1
1–3	3	3	2	1
2–3	5	3	2	2
3–5	1	5	3	3
3–8	1	8	5	3

In cases where a 'medium' value would have been either fractional (eg, 0.5 for the '0–1' range) or required a value not included in the Fibonacci series used for the scoring (eg, 4 for the '3–5' range), the value was rounded down to the nearest number in the series.

A numeric, short integer field, 'Ham_Cnt_N', was created and the numeric values from field 'Ham_Cnt_A' copied into it. For example, a text value of 'H3' in field 'Ham_Cnt_A' has a corresponding numeric value of '3' in 'Ham_Cnt_N'.

The values in fields 'Disp_Scr_A', 'Disp_Scr_N1', 'Disp_Scr_N2' and 'Disp_Scr_N3' for the point which had a hamlet count but no dispersion score in the original data (feature ID 3878) were set to 0. The values in fields 'Ham_Cnt_A' and 'Ham_Cnt_N' for the two points which had dispersion scores but no hamlet counts in the original data (feature IDs 3879 and 3880) were also set to 0.

Exporting the data to shapefile and KMZ

The various polygon, point and line feature classes held in ArcGIS personal geodatabase format were exported to shapefiles and to KMZ files. Because the KMZ format incorporates default symbology into the definition of each feature, it was necessary to set appropriate symbology for the different feature classes. Each feature class was symbolised and then exported to KMZ format using the 'Layer to KML' tool in ArcGIS 9.3.

It was necessary to modify slightly a copy of the 'Terrain Type' feature class before exporting to KMZ. The KML format has a limit on the number of vertices a single polygon feature can contain (see Google Groups KML Developer Support 2007 discussion). As a work-around, one very large polygon classified as 'Lowland', covering the majority of England, was subdivided into three smaller, arbitrary polygons.

ArcGIS 'layer' files (*.lyr) and an ArcGIS map document (compatible with ArcGIS version 9.0 and above) were created utilising the same symbology for the various shapefiles as was used in the KMZ files. The colour schemes for the ArcGIS *.lyr files and the KMZ files

were developed using the ColorBrewer 2.0 website (Brewer *et al* 2009) and the appendix to Cynthia Brewer's *Designing Better Maps* (2005).

CREATING METADATA

Discovery-level metadata compliant with the UK GEMINI version 2.1 standard (Association for Geographic Information 2010) were created in XML format for each shapefile and KMZ file. An XML template was created using the UK Location programme's online Metadata Editor service (Department of Environment Food and Rural Affairs 2010). XML metadata files for each shapefile and KMZ file based on this template were then created and edited using text editing software.

REFERENCES

Association for Geographic Information 2010 *UK GEMINI: Specification for Discovery Metadata for Geospatial Data Resources, v2.1*. London: Association for Geographic Information

Brewer, C A 2005 *Designing Better Maps: A Guide for GIS Users*. Redlands, CA: ESRI Press

Brewer, C A, Harrower, M, Woodruff, A and Heyman, D 2009 *ColorBrewer 2.0: Color advice for cartography* [web page]. Retrieved 5 November 2010 from <http://colorbrewer2.org/>

Department of Environment Food and Rural Affairs 2010 *UK Location Metadata Editor - Beta Release v1.0* [web page]. Department of Environment Food and Rural Affairs. Retrieved 15 November 2010 from <http://locationmetadataeditor.data.gov.uk/geonetwork/srv/en/main.home>

Google Groups KML Developer Support 2007 *Discussions > Getting Started with KML > Vertex Number Limit* [web page]. Retrieved 4 November 2010 from <http://groups.google.com/group/kml-support-getting-started/browse_thread/thread/a1ff189e86ff37a3?pli=1>

Lowerre, A forthcoming 'The Atlas of Rural Settlement in England GIS'. *Landscapes* **12**(1)

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/>

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

Tchoukanski, I 2008a *ET GeoTools for ArcGIS version 9.2 and above* version **9.4** [computer programme]. Retrieved 17 July 2008 from <http://www.ian-ko.com/ET_GeoTools/gt_main.htm>

Tchoukanski, I 2008b *ET GeoWizards for ArcGIS version 9.2 and above* version 9.8 [computer programme]. Retrieved 17 July 2008 from <http://www.ian-ko.com/ET_GeoWizards/gw_main.htm>