EVALUATING UNCERTAINTY IN CLASSIFICATION WITHIN THE
LAND COVER MAP 2000

Thesis submitted for the degree of
Doctor of Philosophy
at the University of Leicester

by

Paul Robinson BSc (Hons)
Department of Geography
University of Leicester

October 2006
To Lucinda and Eve
Evaluating uncertainty in classification within the Land Cover Map 2000

Paul Robinson BSc (Hons)

Abstract

The specific aim of this thesis is to explore the use of object-based metadata regarding data quality within the specific example of attribute uncertainty in the LCM2000. There are three constituent parts to the work. These are using the metadata to identify and describe uncertainty, validating the metadata to ensure that it is informing the user in the way anticipated and exploring what new information can be generated from the metadata.

A review of the state of the art in spatial uncertainty is presented as well as introductions to data quality metadata, land cover mapping and the Land Cover Map 2000 itself. Metadata within Land Cover Map 2000 is explored with respect to identifying, describing and visualising attribute uncertainty. It is demonstrated to be extremely useful and relatively simple to use, allowing comparison between different landscape types and between different geographical areas. The metadata is shown to be valid in that they provide a reliable indication of attribute uncertainty. This is achieved by comparing the map with cumulative evidence from other existing databases that give the extent of particular land cover types. The new information generated from the metadata gives further insight into the uncertainty, providing a simple description of heterogeneity within parcels using indices of dominance. As the metadata is object-based it also allows the spatial units of the map to be broken down and the impact of impurities within the parcel and its neighbour to be examined.

This work provides a strong argument for the inclusion of object-based data quality metadata within digital classified map products, allowing users of the data to assess whether or not the data is fit for the purpose to which they intend to put it.
Acknowledgments

I would like to acknowledge the University of Leicester Geography Department for funding this work and the Centre for Ecology and Hydrology, Ordnance Survey, the National Forest Company and the Forestry Commission for supplying datasets.

Specific thanks go to a number of people. Andy Brewer and Steve Wilkinson for their help with accessing and using data from the NBN. Annette McGrath for her assistance with accessing and interpreting National Forest data. Also Geoff Smith of CEH for all of his considerable help and support. Further thanks must go to the Joint Nature Conservation Committee for support with printing and time to complete this thesis.

Mention has to be made of the gang at Leicester, particularly Jen and Lisa, who made the journey fun along the way.

Huge thanks has to go to Pete Fisher for inspiring this work, for not rolling his eyes all the times I knocked on his office door and for the many hours spent discussing the work. He has ensured that I have learned a huge amount along the way and kept smiling.

Last, and most certainly not least, the biggest thank you has to go to my wife Lucinda who has been there all along, providing support (emotional and financial) and for giving me the occasional push to make sure I made it to the end.
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>iii</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>iv</td>
</tr>
<tr>
<td>List of Figures</td>
<td>x</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xiv</td>
</tr>
<tr>
<td><strong>1</strong> Evaluating uncertainty within the Land Cover Map 2000 – an introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Introduction to spatial data quality and metadata</td>
<td>4</td>
</tr>
<tr>
<td>1.3 The National Forest</td>
<td>6</td>
</tr>
<tr>
<td>1.4 Aims and objectives</td>
<td>11</td>
</tr>
<tr>
<td>1.5 Form of thesis</td>
<td>12</td>
</tr>
<tr>
<td><strong>2</strong> Literature review – uncertainty, data quality and land cover mapping</td>
<td>15</td>
</tr>
<tr>
<td>2.1 Uncertainty in Geographical Information</td>
<td>15</td>
</tr>
<tr>
<td>2.1.1 Types of uncertainty in Geographical Information</td>
<td>16</td>
</tr>
<tr>
<td>2.1.2 Methods for analysing categorical uncertainty</td>
<td>19</td>
</tr>
<tr>
<td>2.1.2.1 The Error Matrix</td>
<td>19</td>
</tr>
<tr>
<td>2.1.2.2 Modelling alternative realisations</td>
<td>21</td>
</tr>
<tr>
<td>2.1.2.3 Geostatistical techniques</td>
<td>22</td>
</tr>
<tr>
<td>2.1.2.4 Alternative set theories</td>
<td>26</td>
</tr>
<tr>
<td>2.2 Data quality and standards - metadata</td>
<td>27</td>
</tr>
<tr>
<td>2.3 Land Cover Mapping</td>
<td>30</td>
</tr>
<tr>
<td>2.3.1 Land Cover Map 2000</td>
<td>34</td>
</tr>
<tr>
<td>2.3.2 Metadata in the Land Cover Map 2000</td>
<td>35</td>
</tr>
<tr>
<td>2.4 Theoretical framework of categorical data</td>
<td>41</td>
</tr>
<tr>
<td>2.4.1 Categorical data</td>
<td>41</td>
</tr>
<tr>
<td>2.4.2 Time, space and attribute</td>
<td>44</td>
</tr>
<tr>
<td>2.5 Consequences of uncertainty</td>
<td>44</td>
</tr>
<tr>
<td>2.6 Identification, description, visualisation and modelling</td>
<td>47</td>
</tr>
<tr>
<td>2.7 Aims and objectives revisited</td>
<td>48</td>
</tr>
</tbody>
</table>
## Contents

### 3 Exploration of uncertainty in Land Cover Map 2000 metadata

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Introduction</td>
<td>50</td>
</tr>
<tr>
<td>3.1.1 Reporting uncertainty</td>
<td>53</td>
</tr>
<tr>
<td>3.1.2 Visualisation techniques</td>
<td>53</td>
</tr>
<tr>
<td>3.2 Methods</td>
<td>55</td>
</tr>
<tr>
<td>3.3 Results</td>
<td>56</td>
</tr>
<tr>
<td>3.3.1 The use of descriptive statistics in assessing uncertainty</td>
<td>56</td>
</tr>
<tr>
<td>3.3.1.1 Frequency distributions</td>
<td>59</td>
</tr>
<tr>
<td>3.3.1.2 Knowledge based corrections by land cover type</td>
<td>64</td>
</tr>
<tr>
<td>3.3.2 Support for the parcel classification</td>
<td>67</td>
</tr>
<tr>
<td>3.3.3 Visualisation of metadata and Knowledge Based Corrections</td>
<td>71</td>
</tr>
<tr>
<td>3.3.4 Visualisation of the parcels where the class is supported by the PPL</td>
<td>77</td>
</tr>
<tr>
<td>3.3.5 Simulation</td>
<td>79</td>
</tr>
<tr>
<td>3.4 Discussion</td>
<td>80</td>
</tr>
<tr>
<td>3.4.1 Descriptive statistics</td>
<td>80</td>
</tr>
<tr>
<td>3.4.1.1 Frequency distributions</td>
<td>81</td>
</tr>
<tr>
<td>3.4.1.2 Knowledge based corrections by land cover type</td>
<td>81</td>
</tr>
<tr>
<td>3.4.2 Support for the parcel classification</td>
<td>82</td>
</tr>
<tr>
<td>3.4.3 Visualisation of per pixel classification metadata and KBCs</td>
<td>83</td>
</tr>
<tr>
<td>3.4.4 Visualisation of support for final classification</td>
<td>88</td>
</tr>
<tr>
<td>3.4.5 Simulation</td>
<td>88</td>
</tr>
<tr>
<td>3.5 Conclusion</td>
<td>90</td>
</tr>
</tbody>
</table>

### 4 Cumulative evidence and supervaluation

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Introduction</td>
<td>92</td>
</tr>
<tr>
<td>4.1.1 Supervaluation and cumulative evidence</td>
<td>92</td>
</tr>
<tr>
<td>4.2 Input databases for cumulative evidence analysis</td>
<td>95</td>
</tr>
<tr>
<td>4.3 Method</td>
<td>99</td>
</tr>
<tr>
<td>4.4 Results of comparison of cumulative evidence analysis with metadata</td>
<td>102</td>
</tr>
<tr>
<td>4.4.1 Metadata from per-pixel classification - woodland</td>
<td>102</td>
</tr>
<tr>
<td>4.4.2 Metadata from per-pixel classification - urban areas</td>
<td>104</td>
</tr>
<tr>
<td>4.4.3 Metadata from per-pixel classification - open water</td>
<td>106</td>
</tr>
<tr>
<td>4.4.4 Metadata from per-parcel classification – woodland, urban and open water</td>
<td>108</td>
</tr>
<tr>
<td>4.5 Results from the analysis of the extent of woodland in the National Forest</td>
<td>110</td>
</tr>
</tbody>
</table>
5 Evaluating uncertainty in classification using indices of dominance

5.1 Introduction

5.1.1 Background

5.1.2 Indices of heterogeneity or dominance

5.2 Methods

5.2.1 Indices of dominance

5.2.2 Spatial autocorrelation

5.2.3 Comparison with biodiversity

5.3 Results

5.3.1 Alternative indices

5.3.2 Variation of dominance across land cover types and levels of classification

5.3.3 Spatial variation

5.3.4 Potential relationship with biodiversity

5.4 Discussion

5.4.1 Alternative indices

5.4.2 Variation of dominance across land cover types and levels of classification

5.4.3 Spatial variation

5.4.4 Potential relationship with biodiversity

5.5 Conclusion

6 Attribute measurement within fixed land parcels and grown regions

6.1 Introduction

6.2 LCM2000 metadata within theoretical framework
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.3</td>
<td>Methods</td>
<td>162</td>
</tr>
<tr>
<td>6.3.1</td>
<td>Definitions</td>
<td>162</td>
</tr>
<tr>
<td>6.3.1.1</td>
<td>Woodland and multiple land cover types</td>
<td>162</td>
</tr>
<tr>
<td>6.3.1.2</td>
<td>Urban land cover types</td>
<td>163</td>
</tr>
<tr>
<td>6.3.2</td>
<td>Parcels with conflicting Per Pixel List and parcel classification</td>
<td>166</td>
</tr>
<tr>
<td>6.3.3</td>
<td>Region growing methodology</td>
<td>167</td>
</tr>
<tr>
<td>6.3.4</td>
<td>Relationship between areas identified as potential increase in woodland and areas under management by the National Forest Company</td>
<td>172</td>
</tr>
<tr>
<td>6.3.5</td>
<td>Multiple classes methods</td>
<td>173</td>
</tr>
<tr>
<td>6.4</td>
<td>Results</td>
<td>175</td>
</tr>
<tr>
<td>6.4.1</td>
<td>Woodland land cover extent</td>
<td>176</td>
</tr>
<tr>
<td>6.4.2</td>
<td>Discord between woodland classification and the per pixel metadata</td>
<td>178</td>
</tr>
<tr>
<td>6.4.3</td>
<td>Urban land cover extent</td>
<td>182</td>
</tr>
<tr>
<td>6.4.4</td>
<td>Discord between urban classification and the per pixel metadata</td>
<td>185</td>
</tr>
<tr>
<td>6.5</td>
<td>Region growing results</td>
<td>189</td>
</tr>
<tr>
<td>6.5.1</td>
<td>Woodland</td>
<td>189</td>
</tr>
<tr>
<td>6.5.2</td>
<td>Urban</td>
<td>192</td>
</tr>
<tr>
<td>6.6</td>
<td>Relationship between areas identified as potential increase in woodland and areas under management by the National Forest Company</td>
<td>195</td>
</tr>
<tr>
<td>6.7</td>
<td>Multiple classes results</td>
<td>198</td>
</tr>
<tr>
<td>6.8</td>
<td>Discussion</td>
<td>202</td>
</tr>
<tr>
<td>6.8.1</td>
<td>Land cover extent by different measures</td>
<td>202</td>
</tr>
<tr>
<td>6.8.2</td>
<td>Discord between the parcel classification and the per pixel metadata</td>
<td>203</td>
</tr>
<tr>
<td>6.8.3</td>
<td>Region growing</td>
<td>205</td>
</tr>
<tr>
<td>6.8.4</td>
<td>Comparison with management areas</td>
<td>207</td>
</tr>
<tr>
<td>6.8.5</td>
<td>Multiple classes</td>
<td>208</td>
</tr>
<tr>
<td>6.9</td>
<td>Conclusion</td>
<td>209</td>
</tr>
<tr>
<td>6.10</td>
<td>Further work</td>
<td>209</td>
</tr>
<tr>
<td>7</td>
<td>Uncertainty metadata in the Land Cover Map 2000 – conclusions</td>
<td>210</td>
</tr>
<tr>
<td>7.1</td>
<td>Aims and objectives of the thesis</td>
<td>210</td>
</tr>
<tr>
<td>7.1.1</td>
<td>Identification, description and modelling, visualisation and prediction</td>
<td>210</td>
</tr>
<tr>
<td>7.1.1.1</td>
<td>Exploration of the metadata</td>
<td>211</td>
</tr>
<tr>
<td>7.1.1.2</td>
<td>Visualisation of the metadata</td>
<td>212</td>
</tr>
<tr>
<td>7.1.1.3</td>
<td>Validation of LCM2000 metadata</td>
<td>212</td>
</tr>
<tr>
<td>7.1.1.4</td>
<td>Generation of new information</td>
<td>212</td>
</tr>
</tbody>
</table>
7.2 Further work
7.3 Final thoughts

References

Appendices

Appendix 1 Definitions of broad habitats as used in LCM2000
Appendix 2 Colour scheme for LCM2000 land cover subclasses
Appendix 3 Simulation of uncertainty based on LCM2000 metadata - Alternative realizations of the Derwent and Colsterworth squares and animated GIF of the Colsterworth square
Appendix 4 Glossary of abbreviations
# List of figures

<table>
<thead>
<tr>
<th>Chapter 1</th>
<th>Evaluating uncertainty within the Land Cover Map 2000 – an introduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Example showing the parcel-based, vector format of LCM2000</td>
</tr>
<tr>
<td>1.2</td>
<td>Location of the National Forest in the midlands of England</td>
</tr>
<tr>
<td>1.3</td>
<td>Topographic map of the National Forest, showing the Landscape Regions designated by the National Forest Company</td>
</tr>
<tr>
<td>1.4</td>
<td>Extract of the LCM2000 showing the National Forest</td>
</tr>
<tr>
<td>1.5</td>
<td>Analysis of the National Forest area as classified by the LCM2000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 2</th>
<th>Literature review – uncertainty, data quality and land cover mapping</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>A conceptual model of uncertainty in spatial data</td>
</tr>
<tr>
<td>2.2</td>
<td>The methodologies by which per parcel and per pixel percentage values are obtained</td>
</tr>
<tr>
<td>2.3</td>
<td>Example of the object level metadata available within LCM2000, giving an explanation of the attributes and what each one tells us about the parcel and it’s classification.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 3</th>
<th>Exploration of uncertainty in Land Cover Map 2000 metadata</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Map showing the location of pilot squares</td>
</tr>
<tr>
<td>3.2</td>
<td>The Derwent square and the Colsterworth square showing the land cover as described in the LCM2000</td>
</tr>
<tr>
<td>3.3</td>
<td>Land cover of the pilot squares as classified by the LCM2000 by area</td>
</tr>
<tr>
<td>3.4</td>
<td>Frequency charts for the PPL percentage values in the Derwent square</td>
</tr>
<tr>
<td>3.5</td>
<td>Frequency charts for the PPL percentage values in the Colsterworth square</td>
</tr>
<tr>
<td>3.6</td>
<td>Frequency charts for the parcel classification spectral probability percentages in the Derwent square</td>
</tr>
<tr>
<td>3.7</td>
<td>Frequency charts for the parcel classification spectral probability percentages in the Colsterworth square</td>
</tr>
<tr>
<td>3.8</td>
<td>The number of parcels affected by different types of KBCs, analysed by the land cover type of the final parcel classification</td>
</tr>
<tr>
<td>3.9</td>
<td>The mean of the highest value in the per-pixel classification analysed by the final class of the parcel (BHSUBVar) - Derwent</td>
</tr>
<tr>
<td>3.10</td>
<td>The mean of the highest value in the per-pixel classification analysed by the final class of the parcel (BHSUBVar) - Colsterworth</td>
</tr>
</tbody>
</table>
List of figures

3.11 Percentage figures from per-pixel classification for each of the five most common land cover types 73
3.12 Percentage spectral probabilities from the parcel classification for each of the five most common land cover types 74
3.13 Locations of parcels with zero values in the PPL 75-76
3.14 Locations of parcels where a probability aggregation KBC has been used 77
3.15 Parcels where the most common pixel class is the same as the final parcel classification at broad habitat level and class variant level 78
3.16 Two potential realisations of the Colsterworth square 79
3.17 Ordnance Survey 1:50,000 map of the area of the Derwent square 85
3.18 Ordnance Survey 1:50,000 map of the area of the Colsterworth square 86

Chapter 4 Cumulative evidence and supervaluation

4.1 Map showing total agreement on the presence or absence of woodland and non-agreement from the cumulative evidence analysis 103
4.2 Graph of per-pixel percentage values for the woodland category from parcels within areas of different cumulative evidence values 104
4.3 Map showing total agreement on the presence or absence of urban areas and non-agreement from the cumulative evidence analysis 105
4.4 Graph of per-pixel percentage values for the urban category from parcels within areas of different cumulative evidence values 106
4.5 Map showing total agreement on the presence or absence of water and non-agreement from the cumulative evidence analysis 107
4.6 Graph of per-pixel percentage values for the open water category from parcels within areas of different cumulative evidence values 108
4.7 Per parcel spectral probability percentages for parcels within areas of different values from the woodland cumulative evidence analysis 109
4.8 Per parcel spectral probability percentages for parcels within areas of different values from the urban cumulative evidence analysis 109
4.9 Per parcel spectral probability percentages for parcels within areas of different values from the water cumulative evidence analysis 109
4.10 Cumulative evidence values within Landscape Regions 112
4.11 Percentage of Landscape Regions allocated to each value from cumulative evidence analysis 112
4.12 Potential range in woodland cover within each landscape region 115
4.13 Landscape regions of the National Forest showing results of cumulative evidence with recently planted areas 116
4.14 Parishes within the National Forest 118
4.15 Percentage of each parish that has a value greater than one from the cumulative evidence analysis for woodland 120-121
Chapter 5  Evaluating uncertainty in classification using indices of dominance

5.1 The National Forest showing the six Landscape Regions within it 129
5.2 Comparison of the dominance calculation from Simpson’s Index with the difference statistic 137
5.3 Comparison of the dominance calculation adjusted from O’Neill with the difference statistic 138
5.4 Comparison of the dominance calculation adjusted from Eastman with the difference statistic 139
5.5 Difference between two highest percentages by land cover class at the broad habitat level 140
5.6 Comparison of the difference between two highest percentages across different levels of classification grouped by broad habitat 141-
5.7 Difference statistic at broad habitat level across the National Forest 144
5.8 Difference statistic at subclass variant level across the National Forest 144
5.9 Broad Habitat level difference statistic and original LCM2000 for the area of the Mease Lowlands Landscape Region 147
5.10 Broad Habitat level difference statistic and original LCM2000 for the area of the Charnwood Landscape Region 148
5.11 Broad Habitat level difference statistic and original LCM2000 for the Needwood and Trent Valley Landscape Regions 149
5.12 Count of recorded species for 1km squares where sample data is available 153
5.13 Mean index values ($I_1$) for 1km squares across the National Forest calculated at the broad habitat level 153

Chapter 6  Attribute measurement within fixed land parcels and grown regions

6.1 Vegetation-Impervious surface-Soil (V-I-S) model 165
6.2 Example of region growing methodology 170
6.3 The extent of woodland within the National Forest as classified within the LCM2000 176
6.4 Parcels in which the per-pixel list woodland values meet the threshold values 177
6.5 Woodland extent as recorded by the 1:200,000 mapping published by Bartholomew 178
6.6 Unsupported woodland parcels classified as woodland by the per parcel classification but that have per-pixel list woodland values less than the threshold values 179
6.7 New woodland parcels that were not classified as woodland by the per parcel classification but that have per-pixel list woodland values greater than the threshold values 180
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.8</td>
<td>The extent of urban areas within the National Forest as classified within the LCM2000</td>
<td>183</td>
</tr>
<tr>
<td>6.9</td>
<td>Parcels in which the per-pixel list urban values meet the threshold values</td>
<td>183-184</td>
</tr>
<tr>
<td>6.10</td>
<td>Unsupported urban parcels classified as urban by the per parcel classification but that have per-pixel list urban values less than the threshold values</td>
<td>186</td>
</tr>
<tr>
<td>6.11</td>
<td>New urban parcels that were not classified as urban by the per parcel classification but that have per-pixel list urban values greater than the threshold values</td>
<td>187</td>
</tr>
<tr>
<td>6.12</td>
<td>Parcels that are not classified as woodland in the LCM2000 but are potentially woodland if merged with their neighbour</td>
<td>191</td>
</tr>
<tr>
<td>6.13</td>
<td>Example of area identified as potential increase in woodland land cover</td>
<td>192</td>
</tr>
<tr>
<td>6.14</td>
<td>Parcels that are not classified as urban in the LCM2000 but are potentially urban if merged with their neighbour</td>
<td>193</td>
</tr>
<tr>
<td>6.15</td>
<td>Example of area identified as potential increase in urban land cover</td>
<td>194</td>
</tr>
<tr>
<td>6.16</td>
<td>Relationship between areas identified as a potential increase in woodland using a 50% canopy threshold criteria and the areas under management by the National Forest Company</td>
<td>196</td>
</tr>
<tr>
<td>6.17</td>
<td>Relationship between areas identified as a potential increase in woodland using a 20% canopy threshold criteria and the areas under management by the National Forest Company</td>
<td>197</td>
</tr>
<tr>
<td>6.18</td>
<td>Location of the single parcel that meets the 75% urban threshold and the 50% woodland threshold when merged with neighbours</td>
<td>199</td>
</tr>
<tr>
<td>6.19</td>
<td>Location of the parcels that meet the 35% urban threshold and the 20% woodland threshold</td>
<td>200</td>
</tr>
<tr>
<td>6.20</td>
<td>Parcel identified as potential increase in both urban and woodland</td>
<td>200</td>
</tr>
</tbody>
</table>
## List of tables

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 1</td>
<td>Evaluating uncertainty within the Land Cover Map 2000 – an introduction</td>
<td></td>
</tr>
<tr>
<td>1.1</td>
<td>Description of the principle characteristics of the different Landscape Regions within the National Forest</td>
<td>9</td>
</tr>
<tr>
<td>Chapter 2</td>
<td>Literature review – uncertainty, data quality and land cover mapping</td>
<td></td>
</tr>
<tr>
<td>2.1</td>
<td>Comparison of classification system used by LCM2000 with broad habitats</td>
<td>36</td>
</tr>
<tr>
<td>2.2</td>
<td>Object-based metadata available with LCM2000</td>
<td>37</td>
</tr>
<tr>
<td>Chapter 3</td>
<td>Exploration of uncertainty in Land Cover Map 2000 metadata</td>
<td></td>
</tr>
<tr>
<td>3.1</td>
<td>Descriptive statistics from the per-pixel list metadata of the pilot squares</td>
<td>58</td>
</tr>
<tr>
<td>3.2</td>
<td>Descriptive statistics from the per-parcel metadata of the pilot squares</td>
<td>58</td>
</tr>
<tr>
<td>3.3</td>
<td>Number of parcels in the two pilot squares in which the most likely class in the metadata is in agreement with the final classification</td>
<td>68</td>
</tr>
<tr>
<td>Chapter 4</td>
<td>Cumulative evidence and supervaluation</td>
<td></td>
</tr>
<tr>
<td>4.1</td>
<td>The sources and characteristics of datasets used as inputs to the cumulative evidence analysis</td>
<td>98</td>
</tr>
<tr>
<td>4.2</td>
<td>Proportion of study area classified as each target land cover type by the datasets used as inputs to the cumulative evidence analysis</td>
<td>99</td>
</tr>
<tr>
<td>4.3</td>
<td>Output of cumulative evidence analysis for woodland, showing the percentage of the National Forest area allocated to each value</td>
<td>102</td>
</tr>
<tr>
<td>4.4</td>
<td>Comparison of Woodland Region dataset with ranked values from cumulative evidence analysis</td>
<td>110</td>
</tr>
<tr>
<td>4.5</td>
<td>Percentage of Landscape Regions allocated to ranked values from cumulative evidence analysis</td>
<td>111</td>
</tr>
<tr>
<td>4.6</td>
<td>Comparison of Woodland Region dataset with ranked values from cumulative evidence analysis, split by landscape region</td>
<td>114</td>
</tr>
<tr>
<td>4.7</td>
<td>Potential range in woodland cover within each landscape region and the effect on that range once woodland region areas are added</td>
<td>114</td>
</tr>
<tr>
<td>4.8</td>
<td>Output of cumulative evidence analysis showing the percentage of each Landscape Region allocated to each value</td>
<td>117</td>
</tr>
<tr>
<td>4.9</td>
<td>Ranked values from cumulative evidence analysis by parish</td>
<td>119</td>
</tr>
</tbody>
</table>
### Chapter 5 Evaluating uncertainty in classification using indices of dominance

5.1 Datasets from which biological samples were obtained via the NBN Gateway 136
5.2 Mean difference values ($I_1$) for the Landscape Regions within the National Forest at different levels of classification 146
5.3 Spatial autocorrelation in difference values ($I_1$) measured using Moran’s I 146
5.4 Index of $I_1$ values across the Landscape Regions compared with the average value for the entire National Forest 150
5.5 Mean Species count data for the landscape regions 152

### Chapter 6 Attribute measurement within fixed land parcels and grown regions

6.1 Analysis of OS Mastermap attribute data for all areas defined as Urban 164
6.2 Criteria to be investigated based on percentage land cover figures 166
6.3 Possible scenarios in region growing analysis 171
6.4 Extent of woodland land cover under different classification criteria 178
6.5 Areas of parcels with discord between the LCM2000 classification and the Per Pixel List - woodland 181
6.6 Per pixel list values for all land cover types from parcels classified as woodland but having conflicting per pixel list values 181
6.7 Extent of urban land cover under different classification criteria 185
6.8 Areas of parcels with discord between the LCM2000 classification and the Per Pixel List - urban 188
6.9 Per pixel list values for all land cover types from parcels classified as urban but having conflicting per pixel list values 188
6.10 Potential increase in woodland areas from region growing analysis 190
6.11 Potential increase in urban areas from region growing analysis 194
6.12 Comparison of potential increase in woodland with areas under management by the National Forest Company 198
6.13 Statistics of the parcels that meet criteria for both urban and woodland 199
“The reader of a map has no choice except to assume each parcel on the map possesses its stated identity, unless otherwise informed. In areas of complex land use patterns, completely accurate delineation of uniform parcels may be impractical. Therefore in such instances the reader must be informed of the presence of impurities within nominally uniform categories by means of statements in the legend or in the report that accompanies each map.”

(Campbell, 1983)

1.1 Introduction

The issue of uncertainty in geographic information has been considered to be of prime importance for some years, which is clear from the fact that it was made a primary research initiative of the National Centre for Geographic Information and Analysis in the US (Guptill, 1989) and aspects of data quality have been included in all geographic data standards since the NCDCDS introduced theirs in the late 1980’s. Data production agencies have often viewed uncertainty as a matter of data quality, but others, particularly in the academic community view uncertainty as a broader issue than that of data quality. While data quality aims to provide metadata on factors such as accuracy and currency of a database, which relates to error (an element of uncertainty) it does not include detail on aspects of uncertainty such as vagueness (See Section 2.1 for further details on uncertainty and Section 2.2 for more on data quality). Issues surrounding uncertainty have become all the more important with the ever increasing use of Geographical Information Systems (GIS) (Burrough and McDonnell, 1998; Longley et al, 1999; Goodchild, 2000a) leading to greater amounts of GIS data being generated and to an ever greater need to be able to exchange information on data quality.
between users along with the data itself. The quote above from Campbell shows that researchers have been thinking for some time that an important element in the way to deal with such problems is for users to be informed and for them to use such information in deciding how such uncertainty is likely to affect their work.

General statements regarding accuracy across a whole map have been successfully used to model uncertainty within the product, thus increasing the information available to the user (Hunter and Goodchild 1995; Goodchild et al, 1992). Statements in the legend, as mentioned by Campbell above, have also been effectively used to model uncertainty (Fisher, 1991b). The ideal scenario however must be to have such information regarding the quality of a product at the level of each individual object (Fisher, 1993). This is becoming increasingly plausible with the increasing processing power and memory available to computer users, as well as the development of techniques and standards for transporting and sharing large amounts of digital data.

The Land Cover Map 2000 (LCM2000) is central to the work described here. It was published by the Centre for Ecology and Hydrology (CEH) in 2001, and is a national database covering all of the UK. The map was created using satellite imagery in a parcel-based, vector format (Figure 1.1), rather than the raster format, which is more commonly used for land cover maps as they are generally derived from satellite imagery. A further innovation for such a national dataset is that details of the production of each land parcel were stored as part of the production process and published in the form of detailed, object-based metadata. This is the first time that such an approach has been taken, to the best of the author’s knowledge, in such a large-scale national product. The metadata published within LCM2000 is therefore novel for such a dataset and is a non-standard type, which makes it worthy of exploration within the context of data quality reporting and of understanding uncertainty within the map.
This chapter introduces the thesis, and in the next section the issues of spatial data quality and metadata are briefly introduced. In the following section the field area used in much of the work reported here, the National Forest in the Midlands of the England is introduced. The final two sections outline the aims and objectives of the thesis and the format that the thesis takes.
"Error laden data, used without consideration of their intrinsic uncertainty, are highly likely to lead to information of dubious value"

(Zhang and Goodchild, 2002; p3).

Information on data quality is important in order for a user to assess the fitness of the data for its intended use (Chrisman, 1984; de Bruin et al, 2001). For a user to be able to assess this they need information about the quality of the data. This data about the data is called metadata, which was defined by Longley et al (2001) as the formalised description of data. A more specific description is provided by Danko (2001) in his notes on the development of ISO metadata standards, which states “metadata allows a producer to describe a dataset fully so that users can understand the assumptions and limitations and evaluate the dataset’s applicability for their intended use”. This actually describes well the metadata provided within the LCM2000, and clearly emphasises the very important point that metadata can inform the user, allowing them to decide if the data is suitable for their needs, just as Campbell described in the quote at the beginning of this chapter. There are some though who would argue that it is a rather limiting definition put forward by Danko, as it clearly states that it is only the producer of the data that can create the metadata. Traditionally this has been the case, but there have been recent calls for new metadata categories to be defined that can be contributed to and disseminated by users of datasets, as they have useful experience that can be drawn upon by other users (Fisher et al, 2004). There are now several standards that have been established, laying out how and what metadata should be reported (e.g. NCDCDS, ISO, ANZLIC, see section 2.2 for further detail).

Metadata can provide information on any aspect of the data and it is not restricted to data held in a digital format, examples being the date of production and scale on a paper map. These
Chapter 1: Evaluating uncertainty within the Land Cover Map 2000 – an introduction

examples are both important for a user to decide if that map is fit for the use they have for that map. For example, the map could be very old, leading to concerns that it is not sufficiently up to date for the user and the scale may not be large enough to provide enough detail at a local scale.

As can be seen, metadata can assist in assessing fitness for use, but more detailed metadata is often required for a user to make specific assessments. For users to really understand the impact of uncertainty on their work they need to create models of the uncertainty and its effects (Goodchild et al, 1992; Zhang and Goodchild, 2002: p193). To carry out this sort of approach they require data about data quality at a more detailed level, which is why some have called for such metadata to be produced as a matter of course (Fisher, 1993a).

The quote at the start of this chapter stresses the importance of informing the reader of a map about its quality. Without further information of some sort, beyond the map itself, users are rendered powerless and must take the map at face value, although this is generally how users treat maps anyway (Hunter and Goodchild, 1995). Suggestions given by Campbell in the quote are giving statements in the legend or in the accompanying report; these will only give information on data quality at the map or dataset level. With recent increases in computing capacity it is increasingly possible to store quality data, as well as attribute data, at the level of the object within the map, which is something that academics have been calling for over a number of years (Guptill, 1989; Fisher, 1993). This is desirable, as it has been shown that uncertainty in maps varies spatially and the use of dataset level information regarding classification accuracy could be misleading when assessing fitness for use, particularly with large databases, such as those depicting land cover at a national level (Longley et al, 2001; p334; McGwire and Fisher, 2001; van Oort et al, 2004).
1.3 The National Forest

Many elements of the work detailed in this thesis utilise the National Forest as a study area. There are a number of reasons for doing this. Primarily this is because it is a specific area designated for particular management purposes as a single unit and therefore detailed analysis of the land cover within the forest would be of interest in relation to the aims and objectives of the company responsible for the forest’s management. It is also of a considerable size (approximately 500 km²), large enough to give good results, while still being a manageable size for the computing resources available to the work. Finally it was an area for which numerous relevant datasets were available.

The National Forest is in the English Midlands (Figure 1.2). It was conceived in 1987 by the Countryside Commission with the intention of dramatically increasing the amount of woodland cover, and so bringing about a range of environmental and social benefits. The first strategy for the forest was published in 1994 and The National Forest Company (NFC) was established in 1995 to deliver that strategy. They have since published a new strategy (National Forest Co., 2004).

On the basis of a landscape assessment carried out in 1994, the NFC has divided the forest into six Landscape Regions, which it effectively uses as management areas in order to structure the plans laid out in the strategy (National Forest Co., 2004). They are designed to represent distinctive landscape character areas within the forest and should allow management, such as new planting projects, to be sympathetic to the existing landscape in terms of scale and design of such schemes (see Figure 1.3). A brief summary of the six regions are given in Table 1.1
Figure 1.2: Location of the National Forest (shaded black) in the midlands of England.
Figure 1.3: Topographic map of the National Forest, showing the Landscape Regions designated by the National Forest Company, which are used as management areas to guide the National Forest Strategy.
Table 1.1: Description of the principle characteristics of the different Landscape Regions within the National Forest (adapted from National Forest Co. Strategy, 2004).

<table>
<thead>
<tr>
<th>Landscape Region</th>
<th>Principle landscape characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calke Uplands</td>
<td>Strong rural character and upland feel. Parklands and two large reservoirs feature in the valleys surrounded by mixed farmland and hedged fields. Plateau has large arable field and sparse woodland. Settlement pattern of villages.</td>
</tr>
<tr>
<td>Charnwood</td>
<td>Rugged, upland character with heathland and rocky knolls. Former Ancient Forest with ancient woodlands as well as parkland and scattered hedgerow trees. Hedged fields on the lower slopes, but stone field walls in the upland areas.</td>
</tr>
<tr>
<td>Mease Lowlands</td>
<td>Rolling agricultural landscape, often with large fields and intact hedgerows. Scattered hedgerow trees and woodlands. Settlement pattern of villages linked by country lanes. Strong rural character.</td>
</tr>
<tr>
<td>Midlands Coalfield</td>
<td>Strong urban influences. Settlement pattern of straggling towns and villages. Areas of new housing and new woodland planting evident. Mining history also evident with areas of opencast coal and clay working and spoil heaps. Agricultural areas open and rolling with small woodlands.</td>
</tr>
<tr>
<td>Needwood</td>
<td>A well wooded landscape with many ancient woodlands, parklands and hedgerow trees. Scattered settlement pattern of villages and hamlets. Mixed farmland with a pattern of hedged fields.</td>
</tr>
<tr>
<td>Trent Valley</td>
<td>Extensive floodplain. Considerable industrial and urban areas as well as sand and gravel workings. Areas of in large agricultural fields. Diverse mosaic of grassland, wetland, scrub woodland and overgrown hedges.</td>
</tr>
</tbody>
</table>
Chapter 1: Uncertainty in land cover mapping – an introduction

Figure 1.4: Extract of the Land Cover Map 2000 showing the National Forest (for key to the colours see Appendix 2)

The LCM2000 picks out the urban areas shown in Figure 1.3, these run down the centre of the National Forest (Figure 1.4, shaded grey). It also shows that these central areas around the main settlements are dominated by agriculture (Figure 1.4, shaded brown), with more mixed areas at either end and in the north. Figure 1.5 shows that improved grassland and arable land cover types occupy over 60% of the area, with urban categories accounting for a further 15%. So while there is a mix of urban and rural areas within the forest, it is a very managed landscape and less than 10% of it is woodland.
Chapter 1: Uncertainty in land cover mapping – an introduction

Figure 1.5: Analysis of the National Forest area as classified by the Land Cover Map 2000 at the subclass level

1.4 Aims and objectives

The aim of this work is to explore the use of metadata regarding data quality within the specific example of attribute uncertainty in the LCM2000. The type of metadata within the LCM2000 is novel and so it is necessary to assess how useful it is in describing uncertainty as well as developing ways in which it can be used.

The principal objective of this thesis is to explore the detailed metadata published in LCM2000, and to assess how much information a user can gain about the map and its fitness for the user’s purpose as described by Campbell (1983) and Danko (2001). In order to achieve this, there are a number of detailed objectives:
• to explore the LCM2000 metadata to see how it can describe variability and uncertainty in the map,
• to assess if the metadata can be used to visualise uncertainty in the map,
• to assess the validity of the metadata as support for the classification, and
• to assess how the metadata can be used to generate new information about the map.

By addressing these detailed objectives it will be possible to demonstrate how useful the metadata is and propose some ways in which it could be utilised.

1.5 Form of thesis

This thesis follows the form of a series of related projects addressing the objectives detailed above, and as such each project contains sections putting the work described within it into the context of the relevant existing literature. There is also a general review of the literature relating to metadata and the analysis of uncertainty in spatial data, which follows this introduction. The projects are then presented as individual chapters and the themes of the whole thesis are then drawn together in the final chapter as outlined below.

Chapter 2 presents a review of the state of the art in spatial uncertainty. It shows the evolution of uncertainty research and relates this to the state of spatial metadata and particularly emphasises data quality. It discusses the relevance of data quality measures to land cover mapping and introduces the object-based metadata in the LCM2000.

Chapter 3 explores the metadata provided in the LCM2000 to describe uncertainty across areas of the map, showing how it can be used to visualise the uncertainty spatially as well as to analyse which land cover categories are more prone to uncertainty than others. It shows
Chapter 1: Uncertainty in land cover mapping – an introduction

how having access to two independent classifications gives greater insight into uncertainty within the map and allows comparison of different areas within it.

Chapter 4 assesses the validity of the per-pixel list metadata by comparing it with alternative datasets in what is termed here a cumulative evidence analysis, based on the theory of supervaluation. This shows that the metadata is capable of adequately describing uncertainty in the map. The usefulness of the cumulative evidence analysis is further investigated by using it to determine the potential extent of woodland in the National Forest, and assessing the impact of management schemes on this extent.

New information is generated from the metadata in Chapter 5 by using indices of dominance to describe the level of heterogeneity within parcels and explore its relationship with biodiversity. The chapter also shows how this facilitates easier analysis, having reduced the data to a single statistic, and can provide new insight as to whether any confusion in classification of the parcel is predominantly between different broad habitats or between variants of one broad habitat. Measures of dominance based on the metadata are shown to follow the general pattern of biodiversity, though a significant relationship could not be demonstrated.

The last of the projects is presented in Chapter 6. Again new information is generated from the metadata, this time by breaking down the spatial units of the map, the parcels, using a region growing methodology. This examines the impact of impurities within the parcel and its neighbour. This underlines the importance of providing quality metadata at an object level.
Chapter 1: Uncertainty in land cover mapping — an introduction

The final chapter draws the themes together by revisiting the aims and objectives detailed above and providing some final thoughts on the work.
Chapter 2: Literature review – uncertainty, data quality and land cover mapping

"So geographers in Afric-maps
With savage- pictures fill their gaps
And o'er uninhabitable downs
Place elephants for want of towns."

(Jonathan Swift)

2.1 Uncertainty in Geographical Information

Geographical data relates to features at or near the surface of the earth (Zhang and Goodchild, 2002). Such data can be extremely useful for many purposes, such as navigation, informing the planning process and analysing the distribution of geographical features or environmental factors (de Bruin et al., 2001). It can therefore be instrumental in developing our understanding of the world we live in and in deciding how we utilise the space and resources it offers. Uncertainty in geographical information is inevitable, as it is, by definition, a representation of real world phenomena and such phenomena are intrinsically complex (Goodchild, 2003). Geographical information must give some description of a phenomenon and its spatial location or extent. Zhang and Goodchild (2002) describe this as being in the form of a tuple \((x, G)\), where \(x\) defines location and \(G\) is a property or attribute describing the phenomenon. This is a straightforward format for the digital storage of this information, but the problem arises when the complexity of the real world is acknowledged. Such complexity needs to be reduced to the tuple format to enable its analysis (Goodchild, 2000; Zhang and Goodchild, 2002). The simplification or abstraction process is a major source of uncertainty, either in the way that complex objects are defined in order to represent them, or in the way that their spatial extent is recorded. This has been described as the problem of definition (Taylor, 1982 cited in Fisher, 1999).
It is very easy to ignore the underlying uncertainty and users often do (Heuvelink, 1998). The reason this is so easy to do is that digital data can often give the impression of definitive accuracy, therefore it appears that the data are fit for the purpose of the user. If this assumption is invalid then this will lead to poor quality information and in turn to incorrect advice and ill-informed decision-making (Fisher, 1999). A further consideration regarding the importance of understanding the uncertainty in spatial data is the desired increase in interoperability of geographical information systems (Ahlqvist et al, 2000). With greater interoperability data becomes more transferable and so will be used by greater numbers of users, all of whom will need to understand how the inherent uncertainty of the dataset will affect their use of it.

The quote from Swift at the top of the chapter seems to be poking fun at cartographers and the way in which they would fill gaps in their maps, and therefore their knowledge, with pictures such as animals and strange creatures. Such practices would not be acceptable today when creating a serious map, but it is still the case that knowledge is not complete and how should such gaps in knowledge be represented? If assumptions are made based on the data available or less reliable data is used to fill these gaps it is important that the user is made aware of this through some form of quality reporting.

2.1.1 Types of uncertainty in Geographical Information

The term uncertainty is used to describe a variety of problems that affect the reliability of spatial data such as accuracy, error and precision (Foody, 2003). Foody (2003) however breaks uncertainty down into just two types, being ambiguity and vagueness. A slightly different description is presented by Fisher (1999) in the form of a conceptual model of some
of the components of uncertainty (Figure 2.1). These components include error, which is when data is incorrect, vagueness, brought about by poor class definitions and ambiguity, which can be created by the use of differing classification systems. Two types of ambiguity are further described, being discord, where an object is identified as being a member of different classes simultaneously and non-specificity, where some form of interpretation is required to assign a class (Fisher, 1999). All of these scenarios can lead to doubt in the data, meaning that further information is required by users for them to be able to decide on the suitability of the data for their needs. Fisher’s model is set out to describe what treatments can be used to deal with the different components of uncertainty and so identifies error as a separate component as it can be analysed using methods of probability. Foody presumably saw error as being a subset of vagueness.

![Uncertainty Diagram](image)

**Figure 2.1:** A conceptual model of uncertainty in spatial data (taken from Fisher, 1999)
Accuracy and precision, as mentioned above, are also terms often used in relation to uncertainty, though they are more general terms than those of error, vagueness and ambiguity. Accuracy, in terms of attribute accuracy in categorical maps, refers to the correct assignment of categories to parcels (Campbell, 1983: p33). This can relate to any of the components of uncertainty described above, depending on how well the object is defined. Accuracy is associated with how closely a map represents the world, as we understand it. Precision on the other hand relates to the level of detail in a map or system of classification (Campbell, 1983: p33). An alternative term for this would be semantic resolution as it relates to the resolution of the class definitions. A map can show a very broad level of detail, such as delineating urban areas from rural ones, but this is not a precise distinction, as it gives no further detail about the urban or rural area.

An important element of Fisher’s model (Figure 2.1) not yet discussed, is the issue of how well the individual object and the class are defined. Critically, if the class is not precisely defined it becomes extremely difficult to assess such notions as accuracy or uncertainty. The concept of clear understanding and meaning relates to semantics and the issue of clear definitions in mapping data is often termed semantic accuracy. This is a thorny issue as it impacts on all levels of the conceptual model of uncertainty (Figure 2.1) though some researchers have proposed methods of assessing the problem of poorly defined geographical concepts. Bennett (2001a and 2001b) looks at the issue from the perspective of logic and natural language in attempting to understand the effects of vague definitions on the interpretation of the concept of a forest. He explores how there can be multiple interpretations of the definition of an object leading to a number of different representations of the spatial extent of that object, suggesting that the technique of supervaluation can assess these interpretations, giving detail on the level of certainty that the object exists at each location on the map. Fisher and Wood (1998) and Fisher et al (2004) suggest a fuzzy set
approach in defining the extent of a mountain by assessing the persistence of the concept of a peak when analysing the topography of an area from a DEM. Both approaches recognise that there is no single 'correct' interpretation of such concepts and that alternatives must be identified and brought into the analysis.

Zhang and Goodchild (2002, p134) describe the conventional method of presenting categorical maps as “exhaustive, non-overlapping areal units separated by a framework of boundary lines, where single discrete labels are ascribed to individual areal units, on the assumption of internal homogeneity”. When considering definition, if that of the class is not adequate it can leave problems of which class definition a specific location most closely matches. As we have seen, a location can have only one label, but it may match more than one class definition. Boundaries are a particularly difficult area as the change from one class to another on the ground is rarely a crisp, sudden change as is implied by the clear boundaries separating the non-overlapping polygons on the map. This can be a problem of definition of the individual object or a problem of locational accuracy or resolution in the depiction of it.

2.1.2 Methods for analysing categorical uncertainty

Considerable research effort has been put into the problem of categorical uncertainty over recent years and the following section will detail some of the work that has been carried out.

2.1.2.1 The Error Matrix

A common method of calculating and reporting uncertainty in categorical maps is by creating an error matrix, sometimes called a confusion matrix. This uses a number of sample points where highly accurate data is recorded, generally through field visits or possibly from interpretation of data considered to be more accurate than that used to create the map, such as
aerial photographs, so the classification at that location is known. This more accurate data is then compared with the data generated by the mapping process within a matrix. Conventionally the classes allocated in the map are listed down the rows and those allocated by the point samples are shown across the columns and the cells of the matrix contain a count value. The numbers within the cells on the diagonal of the matrix therefore give the numbers where the point samples are in agreement with the map.

Such a matrix gives users information about accuracy for the whole dataset and it is broken down across the classes and importantly allows for the calculation of a number of statistics. Possible calculations from the matrix include the percent correctly classified (sometimes called overall accuracy), producer's accuracy (omission errors), user's accuracy (commission errors) and the KAPPA statistic. All of these can tell the user something of the accuracy of the dataset. The overall classification accuracy calculated from the error matrix calculates the probability that a randomly chosen location is correctly classified, but does not take into account chance agreement between the derived and reference data and so will always overestimate classification accuracy (Zhang and Goodchild, 2002; p82). Congalton (1991) gives a comprehensive review of the techniques of producing and using an error matrix.

Two major problems with the error matrix are that the statistics derived from it are not spatial, so they do not reflect the variation of uncertainty across the map and that it relates to the entire database only and do not reflect the situation within a subset of the data (McGwire and Fisher, 2001). The latter point is particularly pertinent in the case of the LCM2000 as it is a national database and most users would only be interested in a smaller subset of the data. It would be necessary therefore to know the exact locations of sample points used to test the product's accuracy, so the user could recalculate the statistics for their study area.
number of survey points would have to be greatly increased in order to supply this data for a majority of users though (McGwire and Fisher, 2001).

2.1.2.2 Modelling alternative realisations

The importance of the spatial relationship within the uncertainty is too significant to ignore so many researchers have looked to develop more sophisticated methods than the error matrix. Modelling of various kinds is offering far better results as models can take into account greater complexity within the data. Models can be created to generate any number of equally probable realisations of the data, or a “population of distorted versions of the same reality” (Goodchild et al, 1992). This can be done using a stochastic process and is often called Monte Carlo simulation. Monte Carlo methods can be based on the Gaussian distribution when dealing with scalar measurements and this is the simplest form of the technique. When dealing with categorical data however it would require that parameters can be defined for use in an algorithm and that the results can be compared through the use of some summary statistic. Such simulations can be used in a wide variety of situations to analyse spatial data and have been successfully used to analyse uncertainty in categorical maps (Fisher 1991b). These methods can be used to quantitatively test a hypothesis, as was done by Fisher (1991b), who was able to define parameters based on soil survey reports and to summarise the outputs using land valuations based on changes in the soils as modelled by the algorithm. The information from the soil surveys gave indications of the amount of inclusions of one soil type in areas that were classified as another. Such information is reported in the legend of the map and as such gives a general statement of uncertainty for that class wherever it appears on the map. Gentle (1998) gives a detailed account of Monte Carlo methods and the mathematical explanations of the technique.
Goodchild et al (1992) produced a model that takes the same category level information used by Fisher (1991b) and were able to introduce the effects of spatial dependence or autocorrelation into the simulation process. By including this in the model it is moving away from randomly modelling the uncertainty and produces a more realistic representation of the real world. A further problem is produced however, as the level of spatial dependence used in the calculations has a major impact on the output, so an appropriate level must be found though the authors note it would be possible to allow the level to vary regionally across the map.

2.1.2.3 Geostatistical techniques

The simulation approaches by Fisher (1991b) and Goodchild et al (1992) used algorithms to create alternative realizations of the same data. More recently, geostatistical methods have been suggested to accomplish this aim. Bierkens and Burrough (1993) proposed Indicator Kriging as a method of directly calculating a conditional (posterior) probability distribution function (CPDF) for the study area. This is a non-parametric method and so is suitable for working with categorical data. Their methodology requires that the classes of the map be given a strict order. Reference locations, which have been classified through direct observation, are then allocated an indicator code. Each location is given the value 1 if the class it has been allocated is the threshold class or below, and 0 if it is above the threshold class (hence the reason why there must be a strict order to the classes). Simple Kriging of this indicator data models the CPDF for the study area (Goovaerts, 1997), giving probability values for unobserved locations, which are therefore based on nearby observations.

Once the CPDF has been calculated it can then be used to estimate the uncertainty at all locations. One method of doing this is called sequential indicator simulation (SIS), which is a method of generating any number of realizations, all of which honour the statistics of the
CPDF. It does not assume a normal distribution, so is ideal for working with such categorical data, and directly generates conditional realizations, which reflect the spatial variation within the map (Bierkens and Burrough, 1993). A number of realizations can then be used as inputs to analytical models, in order to assess the impact of the uncertainty in the data on the output of the model.

Steele et al (1998) used Indicator Kriging to interpolate accuracy of classification from the training locations used in the construction of the classification scheme. Accuracy was measured as the proportion of times that training sites were correctly classified when that location was not used to train the classifier, using a drop out method. A contour map of accuracy was derived from the interpolation. Comparison of this contour map with the original image can show which types of areas have low accuracy, leading to theories as to the cause of incorrect classification within the dataset. There are issues, however, with using training locations as the basis of such calculation, such as whether the sampling is truly random and the assumption that the rate of error is independent of the class (McGwire and Fisher, 2001).

Bierkens and Burrough (1993) point out that an advance on their technique would be to include secondary information into the process so that calculations are not just based on the limited, and often sparse, dataset of directly observed or sampled locations. Kyriakidis and Dungan (2000) proposed a method whereby an exhaustive dataset could be combined, as soft data, with the hard, or more definite, directly sampled data. The exhaustive data is derived from the validation of remotely sensed data. This is calculated in the same way as user's accuracy (an estimate of errors of commission), which is a measure of the proportion of locations that have been allocated to the same category in both the image classification (soft data) and the fieldwork (hard data), and is calculated using an error matrix. The use of user's
accuracy in this way is a “probabilistic interpretation of the error matrix” (Kyriakidis and Dungan, 2000). Using this calculation means that the probability for the soft data is entirely land cover class dependent and therefore is stationary, meaning that it does not vary across the study area within each class.

Kyriakidis and Dungan (2000) use Simple Indicator Kriging with varying means to produce the CPDF. With this method the soft probabilities influence the average spatial variability of the hard probabilities. At the hard data points themselves, the calculated posterior probability is zero or one, as these are not altered in any way by the soft data. At all other points the soft probabilities are altered under the influence of the hard data around them; that influence diminishes with distance. At distances beyond the influence of the hard data, as calculated by the variogram, the soft probabilities are unaltered. Again SIS was used to analyse the calculated CPDF.

de Bruin (2000) has also suggested a methodology whereby hard and soft data can be integrated into a geostatistical analysis to assess spatial uncertainty. The soft data used in this case were posterior probability vectors from the image classifier, which means that soft probabilities are not dependent on land cover class, and therefore not stationary.

The approach put forward by de Bruin (2000) uses a different geostatistical technique, that being collocated cokriging. This method, unlike Simple Kriging with varying local means, as used by Kyriakidis and Dungan (2000), accounts for the spatial cross-correlation between hard and soft data. The implication of this is that estimates from this method will potentially be less influenced by sharp local contrasts in the soft data (de Bruin, 2000). Again SIS was used to model spatial uncertainty from the CPDF.
Work by McGwire and Fisher (2001) assessed the amount of soft data that should be included in the analysis. The soft data available was a classification from a remotely sensed image. The authors compared simulations using differing percentages of this soft data, along with hard data points. Comparisons were carried out by assessing similarity with the original classification and by assessing accuracy, based on comparison with a photo-interpreted map, which was taken to be "ground truth". Effects on both assessments were seen to level off beyond around 10% of soft data being used. At this point the similarity with the original classification was approximately 80% and the accuracy was just over 60%, which compares with an accuracy of 64% for the original classification.

Whilst increasing the percentage of soft data used, up to a certain level, increases the similarity with the original classification, it also reduces the patchiness in the differences between the simulation and the original classification. The authors note that this is primarily due to the simulation being carried out at the individual pixel level, and that the only way to mimic patchiness, as seen in the original data set is to reduce the amount of soft data used. Such a reduction, however, also reduces the correspondence with reality, leading to a difficult decision as to what level of soft data to use (McGwire and Fisher, 2001)

All of the geostatistical methods outlined above utilise the indicator approach, though different forms of it. The indicator approach is a non-parametric geostatistical method, and so suitable for handling categorical data, such as land cover maps. Each method outlined had different data available for analysis. While initially only a hard data set of surveyed points was available, later analyses created, or used image-derived, soft data, which was integrated with the hard data. The important common theme between all of the geostatistical methods discussed is that they account for spatial variation in the uncertainty of the data available.
2.1.2.4 Alternative set theories

Whenever we are dealing with a categorical map we are dividing an area up using sets, whether the user of the map is aware of this or not. In a vast majority of cases the sets being used in such a map are Boolean, meaning that each location is designated as being either a member of a particular set or not (Fisher, 2001). Over recent years alternative set theories have been devised and these bring with them the opportunity to view the way we map things in different ways. In fact the issue of set theory is an important consideration in how to view the problem of uncertainty in general. Fisher (2001) gives a summary of some of these different theories, but it is important to consider some of the different concepts here.

Boolean sets have a membership of 0 or 1, with 0 expressing that the object or location is not a member of the set and 1 denoting that it is. Rough sets were proposed by Pawlak (1982) and provide additional possibilities by adopting a three-valued logic and recognise uncertainty in the categorisation. As well as the member or non-member categories there is another option of a boundary class, which is not definitely a member or definitely a non-member. This logic can relate to spatial uncertainty, being literal boundaries between different areas on the ground or can relate to semantic differences between attribute categories being unclear.

Rough sets could possibly be one way to view the problem of nonspecificity in the conceptual model of uncertainty in spatial data (Figure 2.1), where data relating to locations or distinctions between categories lacks specificity. Another element of that model relating to poorly defined objects is vagueness and one suggested method of treating this issue is fuzzy sets (Fisher, 1999), which were first described by Zadeh (1965). Rather than the three possible distinctions offered by rough sets, fuzzy sets allow the description of the degree to which an object or location matches the definition of a class (Fisher, 2001). The degree of membership is described by a value between 0 and 1. This means that under this approach the
object is not necessarily a member of a set at all (unless it has a value of 1), but the variation in fuzzy values across a map can be analysed.

A final set theory that should be considered is called multisets. Whereas rough sets differ from the Boolean approach by realising that an object is not necessarily either a member or a non-member. Multisets differ by realising that any object does not have to be a member of just one set. It is very possible, particularly when dealing with natural phenomena that an object can justifiably be said to meet the definition of more than one set or category (Fisher, 2001). In the case of land cover, an area of low density housing that has large gardens with trees could easily meet a definition of woodland that is based on the amount of tree canopy coverage. The area would also meet any definition of settlement or residential area as it is occupied by housing. Another example in the case of mapping land use would be an area of shops that have residential flats above them. Such areas meet the definition of a residential area and a commercial one. In a Boolean world both of the examples given would have to be given a single category, so one of the possibilities would have to be chosen as the best fit. When considered using the logic of multisets these areas can be assigned membership to all categories for which they meet the definition.

2.2 Data quality and standards - metadata

Spatial data will always contain inherent uncertainties and errors (Burrough and McDonnell, 1998; Longley et al, 2001; Zhang and Goodchild, 2002). There has been a great deal of work carried out in the last decade or more to decide how best to report the quality of digital spatial databases (NCDCDS, 1988; Guptill and Morrison, 1995; Foody, 2002; Joos, 2003). If such reporting is carried out in a constructive way then the user of the data will be empowered to decide if that data is suitable for their needs. Although, in practice, users may well not be
interested in the uncertainties involved (Goodchild, 1995), there is growing pressure on data producers to provide information which is compliant with International Standardisation Organisation (ISO) specifications on the quality of their products. Such information about the composition and construction of the data is called metadata. The widespread use of technologies such as Geographical Information Systems (GIS) has increased the need for quality metadata, as digital spatial databases are not necessarily a final product in themselves, as was the case with a paper map, but an input into analyses carried out by the user (Morrison, 1995; Smith and Fuller, 2002).

Various proposals for standardised metadata have been created by different organisations, with such aims as allowing easier comparison of datasets from different providers, which will form part of the decision a purchaser has to make between competing datasets (Joos, 2003) and increasing compatibility of datasets (Green and Bossomaier, 2002). Such standards often, if not always, include information regarding data quality. The National Committee on Digital Cartographic Data Standards in the United States published an early example of such a standard created for digital spatial data (NCDCDS, 1988). The committee proposed five components for reporting spatial data quality, being lineage, positional accuracy, attribute accuracy, completeness and logical consistency (see Morrison (1995) for detailed descriptions of these components). Later work by the Commission on Spatial Data Quality of the International Cartographic Association proposed two further components: semantic accuracy and temporal information (Guptill and Morrison, 1995).

Metadata standards created since the NCDCDS report have often only required positional accuracy information (Goodchild, 1995), while others have used various combinations of the data quality components outlined above. ANZLIC (Australia and New Zealand Land Information Council) published their standard in 1998 adopting the same components as

As already discussed, compliance with such standards serves the purpose of allowing easy comparison and increasing compatibility between datasets. The standards, however, are principally intended for use at the dataset level, and metadata at this level does not address the problem of quality varying within a large dataset, as is often the case (Smith and Fuller, 2002). This issue may be advanced by the inclusion of metadata at levels other than that of the dataset. Fisher (1993a) discusses the potential benefits of including quality metadata at all levels from the dataset (or mapset) to the level of the individual object, suggesting examples of metadata that would be relevant at each level. Aalders (2002) envisages metadata reporting of quality information relating to homogeneous subsets of the dataset and Gan and Shi (2002) created a system to manage and update object level metadata that complies with FGDC standards for a topographic map series. The ANZLIC standard discusses the potential for object level metadata, though it is not implicitly built into the reporting structure. Data quality metadata is included at the object level within the 1:250,000 topographic map of Australia (Geoscience Australia, 2003). The ISO19115 (metadata) standard does give the opportunity for quality reporting at a level specified by the user, this could be at the object level, a regional level or for the whole dataset, though any level below that of the dataset is optional (ISO, 2000) and the level of reporting has to be specified if it is more detailed than the dataset level (Joos, personal communication, 2003). In the case of land cover mapping, which is the focus of this work, it could be possible and relevant, to include all of the previously mentioned components of spatial data quality as object-based metadata. That is
with the possible exception of semantic accuracy, which deals with the suitability of the
classification model employed within the dataset.

Goodchild (1995, p76), when discussing an example of a classified soil map, suggested, “we
might go so far as... to tag each polygon with an estimate of the proportions of various
classes present within it”. In other words, use object-based metadata to supply the user with
information regarding data quality, something that had been called for in the literature even
earlier than this by Guptill (1989). This is a step towards recording a probability for all
possible classes at all pixels in a land cover map, rather than a single, inflexible land cover
class (Foody, 1996; Mather, 1999). By using this approach the user is supplied with adequate
metadata to decide whether or not the data is appropriate for their needs in specific terms,
such as for the specific study area in which the user is interested, rather than just in general
terms as with database level quality statements. Database level metadata regarding quality is
still potentially useful in some circumstances such as for semantic accuracy, although this
could be reported at the category level. Reporting quality data at the object level therefore
solves some of the problems of dataset level reporting raised in Section 2.1.2.1. Database
level quality statements have been successfully used to model issues of accuracy in DEM’s
called for distributed models of error and meaningful object level metadata can provide such a
distributed model of quality, which should lead to more spatially responsive models.

2.3 Land Cover Mapping

Land cover and land use maps form a very useful resource for the planning of land
management, new developments and resource use at a variety of scales from local to national
and even global. As well as being a management tool they are widely used in academic
Chapter 2: Literature review – uncertainty, data quality and land cover mapping

research across a variety of disciplines, such as environmental science and hydrology. The LCM2000 is the latest of a number of national surveys of land use or land cover in this country. Campbell (1983) points to the Domesday Book, ordered by William I of England in 1085 as probably the earliest survey of land use on a national scale. It was produced in order to assess resources across the country for the purpose of taxation, but it was not used to produce maps. Thomas Milne created an early map of the land use of London and its surrounding areas in the 1790’s. Though this was not a national map, it was displayed in a way very similar to much later maps, using colour washes and letter codes to represent land use, with these being drawn onto a detailed Ordnance Survey base map (Barber, 2005). An early project to produce a national map of land use for Great Britain was directed by L. Dudley Stamp in the 1930’s and was called the Land Utilization Survey (Stamp, 1948). It was an enormous logistical undertaking requiring thousands of volunteer surveyors, but most of the fieldwork was carried out over the two years of 1930-31 (Campbell, 1983). The outputs of the project included 140 map sheets based on the Ordnance Survey 1 inch (1:63,360) maps with coloured overlay representing land use, very much like Milne’s map.

A further survey, again using predominantly volunteer surveyors, was carried out for England and Wales in the early 1960’s. This survey was directed by Alice Coleman, and while being more detailed than its predecessor, was very similar in approach and again produced maps of the whole area based on field survey carried out over a short period of time (Coleman, 1961). This second survey, carried out in the prosperous period of the 1960’s was designed to contrast with the earlier survey carried out in the depression of the inter-war period, enabling researchers to compare the effects of these different economic climates on the landscape of Britain.
Neither of these early national surveys made significant use of interpreted aerial photography. Such techniques played a very important role in both the World Land Use Survey, which was first proposed by the International Geographical Congress in 1949, and Marschner's survey of Land Use and its Patterns in the United States, published in 1959 (Campbell, 1983). These broad scale mapping projects utilised the potential of aerial photography to capture large amounts of information relatively quickly in comparison with field survey. The launch of the first civilian earth observation satellite, Landsat 1, in the early 1970's provided a further step forward in our ability to collect large volumes of data about the earth's surface. While satellite imagery has a lower resolution than aerial photography, and the cost of producing a satellite and putting it in to orbit is high, it has the benefit that it is continually collecting data once it is there. So the data is collected in a way that is consistent and relatively frequent, making it possible to review data for an area over time. To do the same with aerial photography would require repeatedly chartering flights over specific areas and much smaller areas could be covered in the same timescale.

A further development over the time of the surveys that have been discussed here, apart from the improved technology, is the way that the categories are organised. It was not until the 1970's and the United States Geological Survey land cover mapping program that a fully hierarchical classification system was used for maps based on remote sensing in such a large scale project (Anderson et al, 1976; Campbell, 1983). The LCM2000 uses a classification system based on broad habitats (or target classes), subclasses and variants, giving increasing levels of detail at each level of classification. Earlier surveys used a single level of detail in the main, with just some of the categories having subclasses. An example of this is Stamp's, which had seven categories, but only that of forest and woodland was split into further detail, giving high forest (specified as coniferous, deciduous or mixed), coppice, scrub and forest cut and not replaced (Campbell, 1983).
Typically land cover mapping has involved classifying all points of the target area, whether this be a parcel or an individual pixel, into one of a number of predefined categories. This generalisation of the earth's surface is useful and often necessary to allow analysis, but it necessarily involves loss of detail so it is crucial that it be done in a controlled way (Ahlqvist et al, 2000). A map showing a region using this type of classification has been called an area-class map (Goodchild, 2003; Mark and Csillag, 1989), and is displaying categorical data (see Section 2.4 for further discussion of categorical data).

So far reference has been made to land cover and land use, so it is worthwhile clarifying these concepts by making the distinction between them, particularly as there has been increasing confusion between the terms over recent years (Fisher et al, 2005). While both concepts are mapped in a similar way, producing maps that look very alike, and can actually have similarly named categories, they are philosophically very different. Land cover relates to the physical covering of the earth's surface, whether that is natural or managed vegetation, or man-made features. Land use on the other hand relates to how the land is used by humans, so it has a socio-economic emphasis and will use categories such as industrial or residential rather than built-up or urban fabric, which would be a land cover classes (Brown et al, 2002; Campbell, 1983, pp8-9; Fisher et al, 2005).

Traditionally land cover maps have followed the strict cartographic model whereby all areas of the map are assigned a single category from a fixed classification scheme. A region is therefore represented by a series of non-overlapping, homogeneous objects separated by crisp boundaries that represent immediate change from one category to another (Zhang and Goodchild, 2002; p 134). Such a strict cartographic model does not allow the variability in the landscape to be represented in the map (Campbell, 1983; p32), it also raises questions
about the positioning of the boundaries, as such a crisp delineation between land cover types is often unrealistic, particularly in areas of semi-natural vegetation. A further question is raised by Comber et al (2005) concerning whether users of land cover maps have sufficient understanding of how the land cover categories are actually defined in order to understand what it is that is being mapped. Or, more importantly, if the users are given sufficient information to enable them to understand exactly what the producer of the map has actually mapped. This latter point is an important consideration for users of land cover data, but is not an issue addressed in this thesis.

2.3.1 Land Cover Map 2000

The LCM2000 (see Fuller et al, 2002b and Section 1.1) formed part of the Countryside Survey 2000 (Haines-Young et al, 2000) and aimed to provide a national coverage of widespread broad habitats. These broad habitats are defined under the UK Biodiversity Action Plan (DoE, 1994) and provide part of the basis under which the UK meets its obligations under the UN Convention on Biological Diversity (1992). So national policy was a major driving force behind the adoption of the classification scheme used in the map. Comber et al (2002) argue that there were therefore two major influences on the production process of LCM2000, being the desire for improved science in the production of the map and a policy driven desire to create a map that would assist in the monitoring of international obligations in the management of biodiversity.

The requirement to map broad habitats, which are defined as tools for policy obviously raises technical problems when trying to create a land cover map from satellite imagery. Some broad habitats are too rare or fragmented to be detectable in the imagery or cannot be discerned from satellite images and contextual information with any reliability (Fuller et al,
Chapter 2: Literature review – uncertainty, data quality and land cover mapping

2002b). Problems such as these lead to the 20 relevant terrestrial broad habitats being broken down to 16 Target classes that effectively represent the best matches that could be made to the broad habitats given the technical difficulties. The 16 target classes are further broken down in a hierarchical classification structure into 25 subclasses and 71 class variants (Fuller et al, 2002b; Table 2.1). Although the target classes are not the same as the broad habitats, the terms will generally be used interchangeably here when referring to features in the LCM2000.

2.3.2 Metadata in the Land Cover Map 2000

The techniques and methods used in the production of the map in a vector format (see Section 1.1) were generally seen as superior to the grid based approach of its predecessor the Land Cover Map of Great Britain (LCMGB), particularly in managed areas (Dean and Smith, 2003; Fisher et al, 2002). Vector land parcels were created by a segmentation algorithm, which identified spectrally uniform areas within the image. The algorithm used a region growing procedure from seed points having already identified edge features to ensure that seed points were not located on boundaries (Fuller et al, 2002b). Within the database, users may acquire a considerable amount of object-based metadata, including data quality information, giving detail as to the production history of each parcel (see Table 2.2 for a description of some of the metadata available). To the author’s knowledge, the LCM2000 is the first national spatial database to contain such a level of object-based metadata especially relating to attribute uncertainty.
Table 2.1: Comparison of classification system used by LCM2000 with broad habitats (adapted from Fuller et al., 2002b)

<table>
<thead>
<tr>
<th>Widespread broad Habitat</th>
<th>Target class</th>
<th>Class variants</th>
<th>Number of Variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inshore sublittoral sediment</td>
<td>Sea/Estuary</td>
<td>Sea/Estuary</td>
<td>1</td>
</tr>
<tr>
<td>Standing water/canals</td>
<td>Water (inland)</td>
<td>Water (inland)</td>
<td>1</td>
</tr>
<tr>
<td>Littoral rock</td>
<td>Littoral rock and sediment</td>
<td>Littoral rock</td>
<td>2</td>
</tr>
<tr>
<td>Littoral sediment</td>
<td>Littoral sediment</td>
<td>Saltmarsh</td>
<td>2</td>
</tr>
<tr>
<td>Supra -littoral rock</td>
<td>Supra -littoral rock and sediment</td>
<td>Supra -littoral rock</td>
<td>1</td>
</tr>
<tr>
<td>Supra -littoral sediment</td>
<td>Supra -littoral sediment</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Bogs</td>
<td>Bogs</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Dwarf shrub heath</td>
<td>Dwarf shrub heath</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Montane habitats</td>
<td>Montane habitats</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Broad-leaved woodland</td>
<td>Broad-leaf wood (deciduous/evergreen)</td>
<td>Broad-leaf wood</td>
<td>4</td>
</tr>
<tr>
<td>Coniferous woodland</td>
<td>Coniferous woodland</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Arable and horticulture</td>
<td>Arable and horticulture</td>
<td>Arable and horticulture</td>
<td>22</td>
</tr>
<tr>
<td>Improved grassland</td>
<td>Improved grassland</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Neutral grassland</td>
<td>Semi-natural &amp; natural grasslands &amp; bracken</td>
<td>Rough neutral grass</td>
<td>1</td>
</tr>
<tr>
<td>Calcareous grass</td>
<td>Calcereous grass</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Acid grassland</td>
<td>Acid grass</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Bracken</td>
<td>Bracken</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Fen, marsh and swamp</td>
<td>Fen, marsh, swamp</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Built-up areas, gardens</td>
<td>Built-up areas, gardens</td>
<td>Suburban/rural developed</td>
<td>1</td>
</tr>
<tr>
<td>Continuous urban</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inland rock</td>
<td>Inland bare ground</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>20 relevant broad habitats</td>
<td>16 Target classes</td>
<td>25 Subclasses</td>
<td>71 variants</td>
</tr>
</tbody>
</table>
Table 2.2: Object-based metadata available with LCM2000 (adapted from the full dataset description supplied with LCM2000 by CEH)

<table>
<thead>
<tr>
<th>Title</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SegID</td>
<td>Unique identifier code for each parcel.</td>
</tr>
<tr>
<td>TotPixels</td>
<td>Total number of pixels in each parcel.</td>
</tr>
<tr>
<td>CorePixels</td>
<td>Number of pixels in core area of the parcel used in per parcel classification.</td>
</tr>
<tr>
<td>BHSUBVar</td>
<td>Hierarchical code detailing the dominant land cover type for each parcel, giving broad habitat, subclass and variant. Produced by maximum likelihood classification of averaged spectral response of core pixels within the parcel.</td>
</tr>
<tr>
<td>BHSUB</td>
<td>Hierarchical code detailing the dominant land cover type for each parcel, giving broad habitat and subclass. From same production process as BHSUBVar.</td>
</tr>
<tr>
<td>OpHistory</td>
<td>Processing history of each parcel including input images, spectral probability from maximum likelihood classification of parcel and the number of KBC rules applied. Includes quality data on lineage, completeness and corrections made relating to logical consistency.</td>
</tr>
<tr>
<td>PerPixList</td>
<td>Top five land cover types by area as identified by individual pixel classification, giving percentage of parcel allocated to each type. Produced by maximum likelihood classification of each individual pixel in the image. Not related to BHSUB or BHSUBVar. Includes quality data on attribute accuracy.</td>
</tr>
</tbody>
</table>

In the production of LCM2000, two classifications were carried out (Figure 2.2). One classified each individual pixel and so is referred to here as the per-pixel classification. The other was carried out on each parcel of land once it had been created via the image segmentation algorithm. To classify the parcel the averaged spectral response of the core pixels was used as input to a maximum likelihood classifier. Pixels around the boundary were not included in the average of the parcel where the parcel was large enough to retain a specified minimum number of pixels after the removal of those on the edge. Using only ‘pure’ core pixels in this classification reduced the impact of edge effects through mixed pixels on the classification (Smith and Fuller, 2002). From this parcel classification the spectral probability for each land parcel is reported in the metadata, being an indication of how close the average spectral response of the parcel is to that of training data. Results of the pixel-based classification within the parcel (PerPixList in Table 2.2) are summarised as the percentage of the parcel that had been attributed to each land cover category. The five categories with the highest percentages were stored and are available within the metadata.
These are termed the per-pixel data (or per-pixel list, PPL) and are completely independent from the classification of the parcel as it was carried out in an entirely separate process. The parcel classification was subjected to a knowledge-based corrections process (KBC), so that certain errors were removed and some fine separations of categories included (e.g. acid, neutral and calcareous grass; Fuller et al, 2002b). The per-pixel data was not affected by the KBC process (Fuller et al, 2002b), so it is a potentially very useful tool for analysing the uncertainty within the database. There are, however, parcels that do not have pixel-based data due to gaps in the satellite imagery available. A full example of the attributes available for a parcel is given in Figure 2.3, showing just how much information about the production process is available.

![Diagram of per-pixel and per-parcel classification](http://www.ceh.ac.uk/data/lcm/lcmleaflet2000/leaflet2.pdf)

**Figure 2.2:** The methodologies by which per parcel and per pixel percentage values are obtained.

Land cover map products created using MLC’s have often merely reported the most likely land cover type. This is wasteful as a considerable amount of potentially useful information is
therefore discarded (Foody et al, 1992). The LCM2000 reports the most likely class and also the percentage probability of that class, as produced by the MLC for each parcel. Inclusion of this statistic as well as information from the independent pixel-based classification means that the user has a large amount of information about every parcel within the database.

<table>
<thead>
<tr>
<th>SegID:</th>
<th>SK145754r1</th>
</tr>
</thead>
<tbody>
<tr>
<td>TotPixels:</td>
<td>41</td>
</tr>
<tr>
<td>CorePixels:</td>
<td>11</td>
</tr>
<tr>
<td>BHSUBVAR:</td>
<td>17.1.1</td>
</tr>
<tr>
<td>BHSUB:</td>
<td>17.1</td>
</tr>
<tr>
<td>WBH:</td>
<td>17</td>
</tr>
<tr>
<td>OpHistory:</td>
<td>35:73:0:1:0:0</td>
</tr>
<tr>
<td>PerPixList:</td>
<td>U_c,27:Au_k,12:Us_b,12:Ud_a,7:Au_w,7</td>
</tr>
</tbody>
</table>

What the metadata tells us about parcel SK145754r1:

The parcel contains 41 pixels, of which 11 were used in the parcel classification.

Final class of 17.1.1 is “Suburban/rural developed”.

PerPixList has 27% Urban residential/commercial (17.2.1), 12% Arable unknown (4.2.10), 12% Suburban/rural developed (17.1.1), 7% Inland bare despoiled (16.1.2) and 7% Arable unknown (4.2.10).

OpHistory has six elements:

- Scene No.: 35 tells us that two Thematic Mapper images were used to classify the parcel taken on 12/05/01 and 21/9/97 for summer and winter respectively.
- Spectral probability: Output from MLC given as a percentage and gives a measure of how close the mean spectral response of the core pixels is to the training data.
- Probability aggregation flag: 0 tells us that no probability aggregation rules was applied in Phase 1 KBC’s.
- Phase 1 KBC rules: the number of phase 1 (scene dependent) rules applied, excluding probability aggregation. In this case one such rule was applied.
- Phase 2 KBC rules: the number of phase 2 rules applied, none in this case.
- Flag for other situations: letter codes indicating other factors affecting the classification. Examples include the parcel being used to train the MLC, a void in the imagery that had to be filled manually or a parcel affected by haze in the imagery. None of these affected this parcel.

Figure 2.3: Example of the object level metadata available within LCM2000, giving an explanation of the attributes and what each one tells us about the parcel and its classification. The parcel shown is one of those representing the University of Leicester campus.
Guptill (1989) outlined some remaining challenges in presenting accuracy data at an object level of detail. One of his major points is to be able to describe heterogeneity within features that are generally displayed as single bounded features with a single class attribute, which means that homogeneity is assumed within that feature. The reporting of per-pixel data in LCM2000 goes some way towards this aim. In presenting his challenge Guptill gives examples of describing the variation of tree density within a woodland stand, which is something that the per-pixel data could give information about. He goes on to suggest that a method is required that would tell the user how one edge of the stand has a gradual transition to brush vegetation whereas another has a very sharp transition defined by a road. In this case the per-pixel data would not be of use as the data is giving a summary of the land parcel and we do not know how the variation described alters spatially across the parcel. In order to do this the whole raster classification would be required, though that is a possible way of reporting this data and doing so would allow users to summarise across the parcels themselves (an algorithm commonly available in GIS packages) where that is required, while also giving full information about the spatial variability of the pixel classification.

Mather (1999) asks the question is a pixel a classifiable object? He points out that there must be a correspondence between the spatial scale of the pixel and the coverage of the features to be classified. This question is equally relevant, if not more so, to the parcels being classified in LCM2000. Has the segmentation algorithm created appropriate parcels to allow the target features to be classified? The algorithm has carried out the discrimination of features, the other step in the classification is identification (Mather, 1999) and that is carried out by the maximum likelihood classifier. By reporting both parcel and pixel classifications it may be possible to draw conclusions about both of these steps.
Chapter 2: Literature review – uncertainty, data quality and land cover mapping

2.4 Theoretical framework of categorical data

2.4.1 Categorical data

Land cover maps are categorical data in that they apply a description to all locations within the map by designating all locations to a predefined category. By convention, all locations are assigned one category and one category only, in other words the categories used are mutually exclusive (Campbell, 1983; O'Sullivan and Unwin, 2003). Categorical data can be considered in more detail using the model of scales of measurement first proposed by Stevens (1946). In his original article, Stevens (1946) suggested four scales of measurement, being nominal, ordinal, interval and ratio and also outlined the sort of statistical techniques it would be permissible to use on data of each scale. This model has been criticised for a number of reasons, though many of these criticisms concentrate on two aspects of the model. Firstly the restrictions applied by Stevens to the analytical methods permissible with each scale are felt by many to be too limiting and run the risk of preventing analysts from carrying out a thorough analysis. Secondly, the four elements of the model are seen as being incomplete or too strict to allow them to be applied to real world situations (Velleman and Wilkinson, 1994; Chrisman, 1997). The latter point is underlined by the fact that Stevens himself added a fifth class, that of log interval, some time after his initial article (O'Sullivan and Unwin, 2003). Neither of these criticisms affect the work reported here as the data used clearly fall into the original scales proposed by Stevens and none of the analyses used contravene the originally stated permissible statistics.

Other authors have suggested further additions to the scales, such as cyclic or directional, which can measure compass direction or longitude. Despite the criticisms levelled at it, Stevens' model has been widely used within social sciences (Chrisman, 1997). Chrisman also points out that the objects studied in geography are less straightforward than in psychology,
the field in which Stevens was working (the object in this case being the experimental subject). Geographical objects are inherently vague (Fisher, 1999; Zhang and Goodchild, 2002, p90), particularly because their natural complexity necessitates some level of abstraction for them to be recorded and analysed (Zhang and Goodchild, 2002, p90). The implication of this is that straightforward mechanisms are not always appropriate.

The scales described by Stevens are generally given in an increasing order of complexity. The nominal scale is the most basic and allows entities to be distinguished from one another (Goodchild, 1995; Longley et al, 2001). All entities are assigned an identifier or class. This could be a unique name or the name of a class to which any number of entities can also belong, with the former being a special case of the latter (Stevens, 1946). The identifiers used, be they unique or classes, are merely descriptive and no order is implied or any distance between categories (O’Sullivan and Unwin, 2003). Therefore the identifiers cannot be mathematically manipulated in any way, as it would be meaningless. If it is possible to place the classes in an order then the data is of an ordinal scale. Nominal and ordinal data can also be called categorical (Velleman and Wilkinson, 1994).

Data that has an interval scale enables differences between categories to be quantified, as well as them simply being ranked. This is the first scale that can be described as quantitative (Stevens, 1946), but it has no true zero point, only one supplied by convention. When a scale does have such a true zero then it is on a ratio scale. The simplest ratio scale is that of number itself, which by its nature has a true zero. Both the interval and ratio scales represent continuous data (Velleman and Wilkinson, 1994).

As mentioned above, land cover data are categorical and in Stevens’ scheme they are nominal as it would be impossible to try and order categories of land cover in any meaningful way.
Chapter 2: Literature review – uncertainty, data quality and land cover mapping

The parcel classification in the LCM2000 is a perfect example of this. Given the traditional cartographic model under which land cover is recorded, users must effectively assume that each parcel is an area of homogeneous land cover (Campbell, 1983; pg32), which is often an unreasonable assumption in the case of land cover. The scale and level of precision used in the production of the map affects the impact of this assumption. Highly general maps covering large areas will have large parcels depicting supposedly homogeneous areas, which are more likely to include areas or patches of other land cover types within them. Precision levels and the number of categories can be increased to combat this problem, but that means producing a more detailed map, which will invariably mean a greater cost. One method to get around this problem is to account for inclusions in the category descriptions, by giving an indication at that level of the likely impurities that will occur within parcels of the given land cover type (Campbell, 1983). Information given at the parcel description level can be used to model the uncertainty in classification. Fisher (1991b) used such data to model soil map-unit inclusions by relating these descriptions to Current Agricultural Use Value as used by the state of Ohio for calculating real estate tax for farmers.

While the parcel classification of the LCM2000 follows this conventional model for land cover mapping, the provision of the data in the PPL effectively gives detail of the likely inclusions of all cover types in the parcel at the object level, rather than the category level as in the earlier example. Under Stevens’ model, PPL is ratio data as any parcel that has 50% woodland represented in the PPL can be said to have double the amount of woodland recorded as that of a parcel with 25% woodland. The values in the PPL however effectively provide a membership function for a nominal category (Chrisman, 1997).
2.4.2 Time, space and attribute

Geographical information can be generally viewed as having three components, time, space and attribute (Chrisman, 1997; Zhang and Goodchild, 2002, p2-3). The LCM2000 fits this model, with the time component being the date of the imagery used in its production, the space component being delineated by the individual parcels and the attribute being primarily the land cover type, but also all of the metadata, assigned to the parcel. Chrisman (1997) details a model of how one of the three components can be measured, though strict control is required of the other two. This model utilises and expands the model proposed by Sinton (1978). In order to measure one component, one of the others must be fixed and the third is used as the control. The process used to create the parcels within the land cover map (Fuller et al, 2002b) fits this model extremely well. Time is fixed as being the date of the imagery as has already been mentioned. Attribute, in this case the spectral response recorded in the imagery, is the control as seed points were used and an algorithm then searched for areas of similar values around that point. Space is therefore the component being measured, as the output of the algorithm is a parcel delineating a space of similar spectral response.

The LCM2000 in its published form has a totally fixed time component, and a space component that is effectively fixed into small sub-units of space, being the parcels. In the process of reporting the PPL for each parcel then, Chrisman’s model is still used, but in this case space is used as the control and attribute is being measured. Time is again the fixed component and the attribute being measured is the output of the per-parcel classification.

2.5 Consequences of uncertainty

Ultimately the consequences of uncertainty in geographical data are that inaccurate or low quality information is produced from it. When this is then used in the decision making process the result is likely to be poor decisions. In the case of geographic information being
used for navigation the result could be taking a wrong turn, which is effectively a poor
decision made because of inaccurate data. Of course, the uncertainty may not impact on the
result of using the data at all. The wrong turn may still lead us in the right general direction,
so there may be no time lost, but if the river crossing is marked as a bridge when it is actually
a ferry, then lost time is likely to be the outcome.

Fisher (1994b) when looking specifically at error in spatial databases draws the conclusion
that it is the use to which the database is being put that determines the acceptable level of
error. If the purpose is to produce an inventory, summarising values over an area then it may
well be the case that low accuracy is perfectly acceptable. It may be when very site-specific
information is required that error can have a considerable impact on the outcome of an
analysis.

This relates directly to the LCM2000 as it is designed to be a national database and to produce
regional and national values for the amount of certain land cover types. For this purpose a
general understanding of uncertainty levels would probably be sufficient, but if the map is to
be used as source data for more specific tasks then a greater level of information about
uncertainty in that location would be required to assess whether it is fit for that purpose.

Geographical data is now generally stored digitally in GIS, which provide a wide variety of
analytical tools that enable comparison of different datasets and often now the modelling of
spatial data. Any analytical technique that uses data containing errors or uncertainty will
propagate those problems through to the results of the analysis. The topic of error
propagation in geographical information is a broad area of study that has been the subject of
considerable work and interest over recent years and Heuvelink (1998) provides an excellent
summary. A detailed discussion of error propagation is beyond the scope of this thesis; it is,
however, an important context for any discussion of uncertainty in geographical information. Any error within data that is then used in some form of analysis, will be affected by the process used in that analysis and so will still be present, in some form, within the result, as the result is a function of the input values (Heuvelink, 1998). An understanding of the errors in the initial data will therefore allow the analyst some understanding of the effect that will have on the results. Such effects can be modelled in order to gain a picture of their impact upon the results.

Without any such consideration of the uncertainty within the initial data it is effectively being taken at face value and assumed to be perfectly accurate and precise. This can never be the case when representing real world phenomena in a generalised and simplified form such as a land cover map. Problems in doing this go beyond the ability to both accurately and precisely collect data into issues of the semantics of what is being mapped (see Section 2.1.1).

Land cover mapping can be extremely important data used as an input into models designed to develop an understanding of such things as water run off and the movement of sediments or pollutants through catchments or modelling future land cover (Fisher et al, 1997). Foody (2003) points out that such environmental models can be highly sensitive to the input data, so that changes to the input data can lead to large changes in the model output. Data on the uncertainty of such inputs and an empirical approach to monitoring its likely effect will allow the model to produce tolerances for the output values. Such highly useful information can only be created with an understanding of uncertainty within the initial data and object level data on the uncertainty allows a more stringent empirical analysis than would uncertainty data at the dataset level. In fact van Oort et al (2004) describe how it is possible that outcomes of error propagation analyses can be misleading when using dataset level descriptions of classification accuracy.
2.6 Identification, description, visualisation and modelling

Zhang and Goodchild (2002, p8) describe how research on uncertainty deals with its identification, description and modelling, visualisation and also with predicting the effects of uncertainty on any analysis carried out, if some knowledge is possessed of uncertainty within the input data. The aims of the work described in this thesis are to assess how the object-based metadata included in the LCM2000 can be used in all of these ways to assist in understanding attribute uncertainty within the database. All of the work described here is concerned with identifying uncertainty using the metadata, but the other aspects of the research into uncertainty, as described above, are addressed by different elements of the work in the following ways. Chapter 3 describes the uncertainty using descriptive statistics and uses different visualisation techniques as methods of further exploring and displaying uncertainty. Chapter 4 assesses the value of the metadata as support of the final classification by comparison with alternative information sources, and uses it to model the extent of woodland as a case study. Chapter 5 furthers the description process using indices of dominance, which could further be used in modelling applications, and Chapter 6 describes a modelling application using the metadata to enrich the database, thus creating new information.

Aspinall and Pearson (1995) explain how uncertainties in categorical maps can be viewed in three different components; these being class identity, class heterogeneity and class boundary location. In the main part the work described here addresses uncertainty in the LCM2000 with respect to attribute uncertainty surrounding the accurate classification of land parcels to land cover types, which is what Aspinall and Pearson refer to as class identity. The other two components are also addressed however, by looking at class heterogeneity within land parcels
Chapter 2: Literature review – uncertainty, data quality and land cover mapping

using indices in Chapter 5 and by breaking down the fixed spatial elements in the LCM2000
to review class boundary location in Chapter 6.

2.7 Aims and objectives revisited

Many aspects of the review presented here reflect the theoretical importance in supplying
object-based metadata regarding data quality. Whether this be to make uncertainty
assessment easier, to provide far more detailed and robust inputs into different methods of
analysis, such as modelling, or simply providing a thorough assessment of quality as part of
the production process. The aim of this thesis (Section 1.4) is to explore the use of the
metadata with specific regard to attribute uncertainty within LCM2000. Having described the
metadata in LCM2000 and seen the novel metadata included, it is possible to be more specific
in stating that it is my aim to examine if there is benefit in providing object-based measures of
heterogeneity in spatial objects, as support for the reporting of object-based metadata. This
will be done by addressing the four detailed objectives given in Section 1.4, which are
described in more detail below.

Exploring the LCM2000 metadata to see how it can describe variability and uncertainty in the
map is addressed in virtually all of the work described in this thesis, though the initial step is
to use descriptive statistics to compare different study areas (see Chapter 3).

Assessing if the metadata can be used to visualise uncertainty in the map, as described in
Section 2.6, is a useful and important tool in the analysis of uncertainty. This will be done by
again using study areas and mapping aspects of the metadata as well as using it to generate
animations (see Chapter 3 and Appendix 3).
Chapter 2: Literature review – uncertainty, data quality and land cover mapping

Assessing the validity of the metadata will be done through comparing it with evidence from alternative data sources (see Chapter 4).

Finally, assessing how the metadata can be used to generate new information about the map will be done using indices to summarise the data within parcels into a single statistic (Chapter 5) and by using the metadata to summarise different spatial units by breaking down those published in LCM2000 (Chapter 6).
Chapter 3: Exploration of uncertainty within LCM2000 metadata

3.1 Introduction

A considerable amount of object level metadata is contained within the LCM2000, detailing the history, or lineage, of the parcel and it's classification, and also giving data from a second, independent classification in the per-pixel list (PPL) (see Section 2.3.2 and PerPixList in Table 2.2). Such metadata relates to data quality, and investigating it is the subject of this chapter, which describes initial exploratory analyses of the metadata using descriptive statistics and visualisation techniques to assess how it can be utilised to assess uncertainty in the map.

In order to review and explore the PPL and other metadata within LCM2000, two geographical areas were selected for investigation. These areas are referred to here as pilot squares and their locations are shown in Figure 3.1. Each pilot square is approximately 8x8km and includes nearly 1800 land parcels. They were deliberately chosen to be of similar size, but in contrasting areas in terms of land cover. The north west square is in the Peak District National Park in the area known as the Dark Peak typified by open upland land cover, dominated by semi-natural vegetation communities. It has steep sided valleys and two reservoirs and will be referred to as the Derwent square, after the river that runs through the eastern part. The south east square is in a rural area on the border between Leicestershire and Lincolnshire. It is an area of lowland agriculture and will be referred to as the Colsterworth square, after its largest settlement (Figures 3.2 and 3.3).

Additional attributes of metadata, beyond those normally contained in the standard LCM2000 product were supplied by Dr. Geoff Smith of CEH. Importantly this included data on the five
most likely classes as given by the per-parcel classification and the percentage probability value associated with each of these classes as calculated by the maximum likelihood classifier used in the production process. This is therefore an equivalent of the PPL for the per-parcel classification and so is reviewed here as a comparison with the PPL, which is a part of the standard product.

Figure 3.1: Map showing the location of pilot squares (dark grey) in relation to the county boundaries of England. Cross-hatching indicates the Peak District National Park. Each square is approximately 8x8km
Figure 3.2: The Derwent square (a) and the Colsterworth square (b) showing the land cover as described in the LCM2000. Each is approximately 8x8km in size. See Appendix 2 for LCM2000 legend.

Figure 3.3: Land cover of the pilot squares as classified by the LCM2000 (BHSubVar) by area (see Appendix 1 for detailed list of land cover types and LCM2000 codes)
3.1.1 Reporting uncertainty

It has been common practice for some time for thematic map producers to provide estimates of the accuracy or uncertainty of the map product. Such statistics were produced for LCM2000 and are published in the final LCM2000 report (Fuller et al., 2002a), which claims that the map has around 85% accuracy in categorising Target Classes, the broadest level of categorisation that is mapped. This figure is calculated through comparison with the field survey data collected in the Countryside Survey 2000 and is adjusted to take into account the necessarily different approaches undertaken due to the different techniques used. Such statistics are often calculated by means of an error matrix, which can give statistics for user’s accuracy, producer’s accuracy and overall classification accuracy (see Congalton, 1991 and Section 2.1.2.1 for further detail), though these statistics relate to the entire dataset and give no indication of how uncertainty varies across the map. Retaining data on the production of each object in a map allows more detailed analysis of how uncertainty varies spatially across the map. This can be done through the use of standard descriptive statistics for different areas of the map or by using spatial statistical techniques.

3.1.2 Visualisation techniques

A number of researchers have suggested that visualisation of uncertainty is a very useful method for communicating object level uncertainty to map users (Fisher, 1994a; Dykes, 1997; Bastin et al., 2002; Lucieer, 2004). Providing users with the information that they can use to visualise uncertainty goes beyond a simple presentation of state. DiBiase (1990) describes the importance of visual perception in enabling researchers to explore data, developing pertinent questions and confirming apparent relationships within that data. Both of these processes take place before any attempt to communicate ideas or findings and can form a crucial aspect of research. Being able to visualise uncertainty in the data strengthens the
researchers ability to confirm relationships, as greater knowledge of the uncertainty can influence whether or not the data can support theories or otherwise.

One proposal of how to present uncertainty data is to enable the user to toggle between a map of the uncertainty and the original map or, as has been commonly used, have a map of uncertainty beside the original map (Wel et al, 1998). The data supplied by the PPL could easily be used as an input to such a system, so long as it is visualised in an appropriate way, such as displaying the percentage of the most common land cover type in each parcel. Fisher (1994a) proposed a system of animation, with duration of the presence of a colour on screen depicting uncertainty, so the colours would change frequently when there is high uncertainty. The animation is based on a simulation of known levels of uncertainty within the map, including accuracy calculated for each class, user or producer accuracies or pixel level accuracy (Fisher, 1994a). This idea is investigated further in terms of the PPL in Section 3.3.5, as the PPL allows the simulation to be based on metadata for each individual parcel to produce a stochastic realisation of the map for an area. One problem with many such visualisation techniques is that interpretation of the data being visualised is generally difficult when there is a large number of classes (Lucieer, 2004). For this reason the simulation created here was not carried out at the class variant level, which has 76 classes, but rather at the broad habitat and subclass levels, which have 20 and 26 classes respectively.

Lucieer (2004) suggests that the logical progression from the visual display of uncertainty information is to allow users to use raw data and quality data to directly interact with the classification algorithm. This would seem to be a step beyond what is possible from the use of the LCM2000 metadata, as the original data and classification algorithms would also be required, but it is still possible for the user to make use of the metadata, through visualising it
along with the original map, in order to inform themselves about the uncertainty in the area of interest, or to use it to inform models of the processes that are of interest to them.

3.2 Methods

In order to explore the metadata in the LCM2000, descriptive statistics are used, enabling the construction of a detailed picture of the relationship between the final map and the metadata. By doing so we can develop a greater understanding of uncertainty within the map. Methods used include reviewing means and ranges of the metadata within the pilot squares along with their frequency distributions and comparing them. These values were then further analysed by land cover types assigned to the parcels in the final classification and assessed to see if the category identified in the metadata supports that given in the final classification. Areas with poor imagery due to problems such as cloud cover have been filled manually or with data from LCMGB. In these areas there is no PPL data available and so the locations of these are reviewed. KBCs have been used in situations where the maximum likelihood probability of the most likely class is less than 50%, where results do not match other contextual evidence in the surrounding areas, where external data masks conflict with the result and to correct known problems (Fuller et al., 2002). The use of such corrections means that there is, by definition, increased uncertainty, so the locations and affected categories are also reviewed.

Various elements of the metadata are visualised by mapping the values within the metadata and also where the metadata is in agreement with the final classification. A simulation has been created for each square to review how well the technique works with the metadata. For each parcel the probability of each cover type in the PPL was normalised to ensure that they all equal 100. A random number was then generated between 0 and 100 for each parcel and the cover type associated with that level of certainty was displayed, following the method of...
Chapter 3: Exploration of uncertainty within LCM2000 metadata

Fisher (1994a). By doing this for every parcel a realisation of the pilot square is generated. Twenty such realisations are then displayed using a standard colour scheme and used to create an animated GIF file, which scrolls through the images. If the technique were to be progressed further it would be an improvement to embed the algorithm in the display software so that it runs at display time as Fisher did. The animation stage has been carried out for the Colsterworth square, but this stage was not carried out for the Derwent square, as there are gaps in the data.

3.3 Results

3.3.1 The use of descriptive statistics in assessing uncertainty

For ease of reference the most common land cover type in the PPL is given the name here of PixA, with the second most common being PixB, and so on to PixE being the fifth land cover type. The percentage value associated with PixA has been termed Pix1 and that associated with PixE is Pix5. Similarly the five most likely land cover types from the per-parcel analysis have been called PolA to PolE with their values being Poll to Pol5.

Tables 3.1 and 3.2 show the differences in PPL values and per-parcel likelihood’s between the two pilot squares. The number of parcels with no data (given in brackets in the second column) is the number of parcels where there was no original satellite imagery, due to problems such as cloud cover. Such voids were filled manually or using data from the LCMGB and so no PPL data is available. There are 375 and 2 parcels respectively for the Derwent and Colsterworth squares. The count values of no data in Tables 3.1 and 3.2 also include the parcels where all the pixels in the parcel are accounted for in fewer than the five land cover types, or where fewer than five classes were assigned a likelihood by the MLC.
It is clear from the PPL figures shown in Table 3.1 that there are considerable differences between the two squares. The number of parcels in each square is deliberately very similar, but the number of parcels with no data displays very different patterns. The Derwent square has 375 parcels (21.59% of the total number) with no data in Pix1. In contrast the Colsterworth square has just two such parcels (less than 0.01% of the total). Subsequent pixel values having no data are parcels where there is no need for PPL data as all of the pixels in the parcel have already been accounted for. The number of parcels with no data increases more slowly in the Derwent square as you progress to subsequent levels of the PPL; by Pix5 there are 153 parcels with no data beyond the number of voids (528 in total). In the Colsterworth square there are 259 such parcels. This shows that the mixture of pixel classes within polygons is smaller in the Colsterworth square.
Table 3.1: Descriptive statistics from the per-pixel list metadata of the pilot squares

<table>
<thead>
<tr>
<th></th>
<th>Number of samples (No data)</th>
<th>Mean value</th>
<th>Minimum value</th>
<th>Maximum value</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Derwent square</strong> (1737 parcels)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pix1</td>
<td>1362 (375)</td>
<td>38.17</td>
<td>9</td>
<td>100</td>
<td>16.46</td>
</tr>
<tr>
<td>Pix2</td>
<td>1359 (378)</td>
<td>19.85</td>
<td>4</td>
<td>48</td>
<td>7.02</td>
</tr>
<tr>
<td>Pix3</td>
<td>1339 (398)</td>
<td>12.57</td>
<td>2</td>
<td>33</td>
<td>4.42</td>
</tr>
<tr>
<td>Pix4</td>
<td>1284 (453)</td>
<td>8.85</td>
<td>1</td>
<td>18</td>
<td>3.29</td>
</tr>
<tr>
<td>Pix5</td>
<td>1209 (528)</td>
<td>6.64</td>
<td>1</td>
<td>17</td>
<td>2.61</td>
</tr>
<tr>
<td><strong>Colsterworth square</strong> (1785 parcels)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pix1</td>
<td>1783 (2)</td>
<td>45.13</td>
<td>11</td>
<td>100</td>
<td>21.05</td>
</tr>
<tr>
<td>Pix2</td>
<td>1774 (11)</td>
<td>18.76</td>
<td>1</td>
<td>48</td>
<td>8.06</td>
</tr>
<tr>
<td>Pix3</td>
<td>1728 (57)</td>
<td>11.00</td>
<td>1</td>
<td>29</td>
<td>4.96</td>
</tr>
<tr>
<td>Pix4</td>
<td>1647 (138)</td>
<td>7.63</td>
<td>1</td>
<td>20</td>
<td>3.57</td>
</tr>
<tr>
<td>Pix5</td>
<td>1524 (261)</td>
<td>5.63</td>
<td>1</td>
<td>18</td>
<td>2.74</td>
</tr>
</tbody>
</table>

Table 3.2: Descriptive statistics from the per-parcel metadata of the pilot squares. (This data is not a part of the standard LCM2000 product and was supplied directly by CEH for this study)

<table>
<thead>
<tr>
<th></th>
<th>Number of samples (No data)</th>
<th>Mean value</th>
<th>Minimum value</th>
<th>Maximum value</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Derwent</strong> (1737 parcels)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pol1</td>
<td>1362 (375)</td>
<td>22.44</td>
<td>7</td>
<td>100</td>
<td>13.88</td>
</tr>
<tr>
<td>Pol2</td>
<td>1357 (380)</td>
<td>14.17</td>
<td>1</td>
<td>47</td>
<td>6.46</td>
</tr>
<tr>
<td>Pol3</td>
<td>1351 (386)</td>
<td>10.09</td>
<td>1</td>
<td>25</td>
<td>3.18</td>
</tr>
<tr>
<td>Pol4</td>
<td>1349 (388)</td>
<td>7.55</td>
<td>1</td>
<td>17</td>
<td>2.09</td>
</tr>
<tr>
<td>Pol5</td>
<td>1261 (476)</td>
<td>6.19</td>
<td>1</td>
<td>12</td>
<td>1.80</td>
</tr>
<tr>
<td><strong>Colsterworth</strong> (1785 parcels)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pol1</td>
<td>1783 (2)</td>
<td>53.17</td>
<td>12</td>
<td>100</td>
<td>23.96</td>
</tr>
<tr>
<td>Pol2</td>
<td>1707 (78)</td>
<td>19.55</td>
<td>1</td>
<td>49</td>
<td>9.34</td>
</tr>
<tr>
<td>Pol3</td>
<td>1574 (211)</td>
<td>10.49</td>
<td>1</td>
<td>29</td>
<td>5.51</td>
</tr>
<tr>
<td>Pol4</td>
<td>1472 (313)</td>
<td>6.73</td>
<td>1</td>
<td>20</td>
<td>3.86</td>
</tr>
<tr>
<td>Pol5</td>
<td>1248 (537)</td>
<td>4.81</td>
<td>1</td>
<td>15</td>
<td>2.89</td>
</tr>
</tbody>
</table>

The ranges of PPL values are very similar between the two squares, as are the standard deviations of the values in each PPL level. Though at the Pix1 level the minimum value in the Colsterworth square is slightly higher, as are the mean and standard deviation. So the values are generally higher but with a slightly wider distribution.
Descriptive statistics for the parcel classification (Table 3.2) show great similarity to those of the PPL. Again there are more parcels with no data at the Pol5 level in the Colsterworth square, despite there being fewer voids. Also the mean and standard deviation of Pol1 values are considerably higher in the southeast, while the values at the other levels are very similar, as was the case with the PPL. In this case though the differences at the Pol1 level are greater than at the Pix1 level, showing that the parcels are classified with substantially higher certainty in the Colsterworth square.

### 3.3.1.1 Frequency distributions

The percentage values from the PPL (Pix1 to Pix5) and the five highest likelihood values from the per-parcel classification (Pol1 to Pol5) were analysed by comparing the frequency histograms from the two pilot squares (Figures 3.4, 3.5, 3.6 and 3.7). The zero values from voids are not shown on the graphs in order to allow easier comparison, but are detailed in Tables 3.1 and 3.2 as the polygons with ‘no data’.

Looking at the PPL values first, the shapes of the frequency histograms are extremely similar between the two squares (Figures 3.4 and 3.5). When reviewed carefully however there are several important differences. The Pix1 graph (the percentage value of the most common land cover type) for the Colsterworth square, while being positively skewed like that of the Derwent square, has considerably more high values. The graphs of all the other levels are almost identical between the squares but the Colsterworth square always has more parcels in the lowest value bar, showing that the values are deceasing as we move through the levels. This can also be seen in Table 3.1, as the mean values of Pix2 to Pix5 are all slightly lower in the Colsterworth square.
The per-parcel values (Figures 3.6 and 3.7) show less similarity between the squares than was evident in the PPL values, though overall patterns are not completely different. Care must be taken when interpreting these graphs as the y-axes are different between the two squares to enable the distributions to be shown clearly.

Figure 3.4: Frequency charts for the PPL percentage values in the Derwent square.
Figure 3.5: Frequency charts for the PPL percentage values in the Colsterworth square.
Chapter 3: Exploration of uncertainty within LCM2000 metadata

Figure 3.6: Frequency charts for the parcel classification spectral probability percentages in the Derwent square.
In the Derwent square there are very few parcels with high values, even at the Pol1 level, which is why a greater number of probability aggregation KBCs were required in this square (see section 3.3.1.2). In the Colsterworth square there are far more parcels with high values and even a spike at the maximum value bar. This is showing that quite a number of parcels have a spectral response very close to just one cover type in the training data. The Colsterworth square in levels Pol2 to Pol5 has consistently more parcels in the lowest value bar as well as more zero values (Table 3.2) showing that in more cases the subsequent levels
have low values and so there is a considerable difference between these and the top level (Pol1).

3.3.1.2 Knowledge based corrections by land cover type

Knowledge based corrections (KBCs) were used in the production of LCM2000 for a number of purposes, such as distinguishing between acid and calcareous grassland (Fuller et al., 2002b). When a KBC was fired this diverged from the standard classification protocol, using such methods acknowledges that the protocol is not able to accurately identify all the target classes, or may at times give uncertain responses, and intervention of some form is required. It is particularly useful then to be aware of how often the general production process is affected in this way, where such corrections have taken place and which land cover types are most affected.

Figure 3.8 shows the final land cover type of all parcels where a KBC was fired, split by the type of KBC. The first thing to note from the three graphs is that the Derwent square is affected by more KBCs than the Colsterworth square. This is potentially misleading as there can be more than one type of KBC fired for each parcel, but the number of parcels affected by a KBC of some type is actually 941 in the Derwent and 272 in the Colsterworth square, representing 54.17% and 15.24% of parcels respectively.

External context, or phase 2 KBCs used external information to provide context to the analysis. Grasslands were identified as acid, neutral or calcareous using a soil mask and this is why there are so many grassland parcels affected (classes 5, 6, 7 and 8 in Figure 3.8c). Bog is the only other type affected, which is because a mask was used to identify where peat was located. Any spectral response relating to relevant vegetation would therefore be converted to
Chapter 3: Exploration of uncertainty within LCM2000 metadata

A bog category in areas where peat occurs and so the categorisation of bog is entirely dependent on external information.

The probability aggregation KBC was designed to review parcels where the likelihood of the most likely class for the parcel was less than 50%. In these cases the other classes identified in the top five were examined to see if together they pointed to a different target class being more likely. When this was the case the second choice class usually took precedence (Fuller et al, 2002b). Figure 3.8a shows that many of these KBCs occurred in parcels that were given the final target classes of 1, 5, 8 and 12, which are woodland, improved grassland, acid grassland and bog respectively. There are a few in arable types (class 4), but it seems that in most cases such confusion was treated by allocating the class 4.10, being arable horticulture of unknown type. Forty arable parcels in total required a probability aggregation KBC out of a total of 972 arable parcels in the Colsterworth square. Suggesting that the classification was not often confusing different arable types. Heath, bog, bare ground and mixed woodland (classes 10, 12, 16 and 1.1.2) required most KBCs of this type. These are highly variable land cover types so it is not surprising that there are a large number of parcels that did not exhibit a spectral response close to that of the training data.

Virtually all the scene dependent, or phase 1 KBCs (Figure 3.8b) related to inland bare ground (class 16.1.1). These KBCs examined surrounding parcels to ensure that there were not unlikely combinations such as arable areas within towns (Fuller et al, 2002b). Categories such as urban or arable have probably been converted to inland bare ground because semi-natural vegetation types like heath surrounded them.
Figure 3.8: The number of parcels affected by different types of KBCs, analysed by the land cover type of the final parcel classification, (a) shows probability aggregation KBCs, (b) shows phase 1 (scene dependent) KBCs and (c) shows phase 2 (external context) KBCs. See Appendix 1 for land cover codes.
Figure 3.8 (cont.): The number of parcels affected by different types of KBCs, analysed by the land cover type of the final parcel classification, (a) shows probability aggregation KBCs, (b) shows phase 1 (scene dependent) KBCs and (c) shows phase 2 (external context) KBCs. See Appendix 1 for land cover codes.

3.3.2 Support for the parcel classification

Another way to examine the PPL is to explore whether it lends support to the final classification (see Table 3.3). Parcels where a data void has been filled cannot be taken into account in this comparison as no PPL is reported for these parcels.

Table 3.3 shows that the most likely land cover type given in the first of the options from the per-parcel list is more often in agreement with the final classification than the per-pixel list (PPL). There is also little difference between the levels of classification (broad habitat or variant) within the per-parcel data, so there are few parcels where the most likely in the parcel list is not the same variant class as the final classification, but that are the same broad habitat. This is an expected result as the final classification is based on the per-parcel analysis. It is more interesting to compare the contrasting stories that emerge from the PPL values in the
two squares. Disregarding the parcels affected by data voids, the Derwent square has less than 33% of remaining parcels where the PPL gives direct support to the final classification. The Colsterworth square on the other hand has almost 78% of parcels with the PPL supporting the classification. These figures drop to 25% and 70% respectively when considering the class variant level.

<table>
<thead>
<tr>
<th>Table 3.3: Number of parcels in the two pilot squares in which the most likely class returned by the per-pixel and the per-parcel classifications is in agreement with the final classification at the broad habitat level and the variant level</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of data voids</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Per Pixel List data</strong></td>
</tr>
<tr>
<td>Broad Habitat</td>
</tr>
<tr>
<td>Number in agreement</td>
</tr>
<tr>
<td>Number not in agreement</td>
</tr>
</tbody>
</table>

This analysis of support for the final classification can be extended to explore what the percentage values from the PPL are across land cover types (Figures 3.9 and 3.10). The Colsterworth square has generally higher percentage values for the most common land cover type in the PPL, with an average value of 45.07%. The average value in the Derwent square is 29.93%. This can be clearly seen when the values are split out across final land cover types, as the Colsterworth square has 12 classes with a mean above 40% (Figure 3.10), whereas the Derwent square has only six classes meeting that value (Figure 3.9).

A further difference between the two squares displayed in Figures 3.9 and 3.10 is that there is a greater difference between the PPL values of parcels where there is support and those where
there is not in the Colsterworth square. The difference in mean values between the parcels with support and those without is 11.35% in the Colsterworth square and only 1.58% in the Derwent square. Meaning that not only are there higher levels of support in the Colsterworth square in terms of numbers of parcels, but it is also greater numerically within the PPL values.

Of course all the figures described here and shown in Figures 3.9 and 3.10 are mean values and so greatly influenced by the sample size, but still reliably show the differences between the two pilot squares.

**Figure 3.9**: Derwent Square: The mean of the highest value in the per-pixel classification (Pix1) analysed by the final class of the parcel (BHSubVar) and further split by whether the class of that pixel (PixA) is in agreement with BHSubVar at the variant level.
**Figure 3.10:** Colsterworth Square: The highest value in the per-pixel classification (Pix1) analysed by the final class of the parcel (BHSUBVar) and further split by whether the class of that pixel (PixA) is in agreement with BHSUBVar at the variant level. See Appendix 1 for land cover codes.
3.3.3 Visualisation of metadata and Knowledge Based Corrections

Section 3.3.1 explored the LCM2000 metadata statistically, this can be very informative enabling a user to explore the metadata for an area and draw new conclusions about the map from the results. All of these methods have given a good representation of each pilot square as a whole or for each land cover type, but have not given the user any information about variability across that area. By visualising the metadata it is possible to start doing just that (Figures 3.11, 3.12 and 3.13), though it is not just uncertainty data that should be treated in this way, any form of spatial data is best conveyed in a visual way, hence the importance of cartography. The fact that spatial structure is presented and understood more easily is one of several factors motivating the use of graphical techniques (Beard and Buttenfield, 1999). Other factors include the fact that the perceptual skills of humans are primarily visual and learned very early in life and that visual methods are very efficient for communicating large amounts of data (Beard and Buttenfield, 1999; Goodchild, 2000b).

The PPL values shown in Figure 3.11 demonstrate what has already been shown in the frequency distributions (Figures 3.4 and 3.5) with very few high values in the Derwent square and the generally higher values, though with a wider distribution, in the Colsterworth square. What can also be seen however is the distribution of where the high and low values occur. It is clear that there is a more uniform spatial distribution of values in the Colsterworth square, though the highest values tend to occur in the south of the square at the Pixl level. The Derwent square shows a very specific pattern with two areas of low or zero values running diagonally from north west to south east. The highest values are generally in the west of the square.

The per-parcel classification likelihoods shown in Figure 3.12 exhibit very similar spatial patterns to the PPL values (Figure 3.11), though the differences between the two squares
appear even greater. Overall impressions given by the maps are again what could be expected from reviewing the frequency distributions in Section 3.3.1.1, with most parcels in the Derwent square having low values, though with this colour scheme the spatial distribution appears more scattered. In the Colsterworth square the progression through the levels shows that most parcels have either one or two land cover types far more likely than any others.
Chapter 3: Exploration of uncertainty within LCM2000 metadata

Figure 3.11: Percentage figures from PPL for each of the five most common land cover types. Colour scheme is equal interval greyscale, with five equal categories and darkness increasing as percentage value increases. No parcels had a value of more than 20 for Pix4 or Pix5 meaning that there are no shaded polygons so these are not shown.
Figure 3.12: Percentage spectral probabilities from the parcel classification for each of the five most common land cover types. Colour scheme is equal interval greyscale, with five equal categories and darkness increasing as percentage value increases. No parcels had a value of more than 20 for Pol4 or Pol5 meaning that there are no shaded polygons so these are not shown.
Figure 3.13: Locations of parcels with zero values in the PPL. Where Pix1=0 a KBC has been applied to fill a data void, otherwise all of the pixels have been accounted for in the more common land cover types (zero value parcels are shaded black).
The location of data voids is extremely important when reviewing the metadata (Figure 3.13 a and b) as these are parcels with no PPL values. It is also of interest to see where zero values occur in later levels of the PPL (Figure 3.13 c to j), as this shows which parcels have fewer land cover types identified within them by the per-pixel classification. Figure 3.13 j shows that there are many such parcels in the south of the Colsterworth square, the area where the most high Pix1 values were to be found.
a) Derwent square  

b) Colsterworth square

**Figure 3.14:** Locations of parcels where a probability aggregation KBC has been used (shaded black) within the Derwent (a) and the Colsterworth (b) pilot squares.

Probability aggregation KBCs are necessary when the maximum likelihood classifier can only give low likelihood matches. There must then be only low confidence in the final category of such parcels. The locations of the parcels where these KBCs were required are mapped in Figure 3.14, which shows that there are many more occurrences in the Derwent square, though no strong pattern is evident in either square.

### 3.3.4 Visualisation of the parcels where the class is supported by the PPL

The numbers of parcels where the PPL is in agreement with the final parcel classification have already been reviewed, along with the percentage values from the PPL in these cases. It is also likely to be of interest to the user, where these parcels are. The spatial pattern of this support could be of great importance to them. Figure 3.15 shows these patterns, though the pattern of data voids (Figure 3.13 a and b) should also be considered when looking at these, particularly the Derwent square (Figure 3.15 a and b).
Chapter 3: Exploration of uncertainty within LCM2000 metadata

Figure 3.15: Parcels where the most common pixel class is the same as the final parcel classification (shaded black) for the Derwent square at broad habitat level (a) and class variant level (b) and for the Colsterworth square at broad habitat level (c) and class variant level (d).

The most obvious difference between the two areas is that the Colsterworth square has a far greater level of support for the classification from the PPL. When viewed at the broad habitat level nearly 85% of the parcels have the most common pixel class (PixA) in agreement with the final parcel classification as opposed to less than 40% in the Derwent square (Table 3.3). There is a smaller difference between the levels in the Derwent square, so only 65 parcels had...
a different class variant of the same broad habitat. Though this is probably related to the fact that much of the Colsterworth square is of the arable broad habitat, which has the highest number of class variant categories within the classification scheme and so the highest potential for confusion.

3.3.5 Simulation

The animations produced from the simulated metadata are attached in a Microsoft Powerpoint file as Appendix 3. They show simulations of the metadata at the broad habitat classification level and at the subclass level. Two examples of the potential realisations are shown in Figure 3.16, which illustrate how the overall impression of the images is similar, but on close inspection there are considerable differences in the detail. Showing just two realisations side by side like this though is not particularly helpful. Such illustrations are far more useful when animated as the viewer can gain a general insight into how much change occurs across an area, or can concentrate on a small area of interest to get a more detailed impression of the uncertainty.

Figure 3.16: Two potential realisations of the Colsterworth square at the subclass level
3.4 Discussion

3.4.1 Descriptive statistics

Analysing the metadata for two distinct areas using descriptive statistics and frequency histograms of the values has given a clear picture of differences in the classification between the areas and given insight into the way it works in different environments. Any study using the LCM2000 and comparing two areas would be able to use this additional information in the comparison, or to inform a model. If a study had a good reference dataset from fieldwork a more detailed analysis could also be carried out comparing those data with the metadata. The pattern emerging from the metadata is that the more managed, arable area around Colsterworth has fewer corrections, slightly higher PPL values and fewer land cover types per parcel being detected in the per-pixel classification. Taken together this suggests that the classification in this area is more reliable than the upland semi-natural area in the Derwent square.

A clear result from the analysis of descriptive statistics for the per-parcel data is that the difference in the PoI1 values between the two squares is much greater than the equivalent difference in PPL values (Tables 3.1 and 3.2). Such a result says a lot about the production process of the map and the way that a mean spectral response is used in the parcel classification. This method is shown to be far more suitable in areas with generally homogeneous cover such as in arable fields than in areas of semi natural vegetation, which have a far more varied spectral response. A similar conclusion was reached by Dean and Smith (2003), who recommended creating a pixel map for semi-natural areas and a parcel one in areas with a more man-made landscape.
3.4.1.1 Frequency distributions

Reviewing the frequency distributions is another way of showing the data in Tables 3.1 and 3.2 in graphic form. The results underline the patterns already identified though they do provide a visual comparison between areas, which is very straightforward and may be a preferable method for some users.

3.4.1.2 Knowledge based corrections by land cover type

Three different types of KBC are identified in the metadata, being probability aggregation, internal context (or phase 1) and external context (or phase 2). When considering these in terms of understanding uncertainty within the map each type needs to be viewed differently. In the case of external context, important issues to consider are the accuracy and resolution of the external datasets used. If they are unreliable then the errors within them will propagate through to the classifications based upon them. Likewise if the resolution is inappropriate to the minimum mappable unit of the final map, then this will lead to incorrect classifications around the boundaries within the external dataset. With internal context KBCs it is important that the rules applied are logical for the area being analysed, so that erroneous classifications do not replace valid ones. Probability aggregation is more related to the uncertainty inevitably created by the protocols used in the classification system. By examining the land cover types effected by these KBCs it becomes clear that those most effected are generally variable land cover types for which it is more difficult to define typical areas in order to train the classifier (Figure 3.8a).

Examining the KBCs is informative and knowing where the probability aggregation rules were applied can give a direct insight into areas and land cover types classified with low certainty. This is, however slightly limiting as the user only knows when a rule has been fired.
due to a threshold being met (less than 50% likelihood). It is far more useful to actually review the figures from the maximum likelihood classifier (the most likely value is in the metadata) or the value derived from the per-pixel classification. By doing this users can see how these values vary across the map and can make their own thresholds if required.

3.4.2 Support for the parcel classification

Describing the values of the metadata gives a useful impression of the uncertainty within the area, but an important question is then whether or not the categories with high values are supporting the final classification or not. A high value for the same class lends considerable confidence to the classification, a low value for the same class gives cause for concern, but a high value for a completely different class shows confusion in the map. Table 3.3 tells us that the Colsterworth square has massively greater support for the final classification from the PPL. Showing that the parcels in that area have a closer relationship with their constituent parts, the pixels within them. When we consider this within the terms of the classes present in the two squares (Figures 3.2 and 3.3) we can see that the Colsterworth square is dominated by arable land cover types, which by their very nature are more homogeneous. The pixel classification is detecting this and the averaged spectral values for the parcel appear to be easier to classify as a consequence of this homogeneity.

The analysis of the level of support offered to the final classification by the PPL clearly shows that the support is not only present in more parcels in the Colsterworth square, but it is also numerically higher in the PPL values. There is a considerable difference in values between the parcels where PixA and the final classification are the same and the parcels where they are not (Figure 3.10). This difference is far less pronounced in the Derwent square (Figure 3.9). It is clear then that in the more managed, homogeneous environment, high PPL values are
giving a clear indication of certainty, a result that says much about the method of using the mean spectral values of the parcel in the classification process. Such a method is far less likely to give clear results in heterogeneous and therefore spectrally variable areas.

3.4.3 Visualisation of per-pixel classification metadata and KBCs

The field of visualisation within geographical information science is a wide and complex one. It is an area in which there has been considerable effort and innovation over recent years (Bastin et al., 2002; Beard and Buttenfield, 1999; Ehlschlaeger et al., 1997; MacEachren and Kraak, 1997; van der Wel et al., 1999). A detailed review of this field is beyond the aims and scope of this work, but it is an important tool that relates to the way in which object level metadata can be explored and utilised.

DiBiase (1990) as cited in MacEachren and Kraak (1997) presented a model of visualisation based upon exploratory data analysis that split visualisation into four stages. The first two stages were exploration and confirmation and are private visual thinking stages. After these come the public visual communication stages of synthesis and presentation. The visualisation elements of the work presented in this chapter relate to the first two stages. Using the LCM2000 metadata in a visual way to assist the user in exploring the map and its suitability for purpose. Examples of the public visual communication stages are presented in Chapters 5 and 6.

As part of the private visual thinking stages, it is useful to compare the data you have with other datasets, so the Ordnance Survey 1:50,000 maps of the two pilot squares are shown in Figures 3.17 and 3.18. They are helpful in understanding the patterns of PPL values described in Section 3.3.3. The two diagonal patches of no value include the reservoir and the
steep hills to the east of it as well as the high area called The Ridge to the north of Alport Moor (Figure 3.17). It is highly likely that these areas were covered with cloud in the imagery. The areas with the higher PPL values tend to be those less steep areas, for example to the north west of Ridge Nether Moor, north and west of Bleaklow Hill and to the east of Cow Stone Edge.

In the Colsterworth square it is more difficult to identify relationships with the PPL values and this scale of mapping, as much of the area is open farmland without many features other than the roads. It is interesting when considering the more uniform spatial distribution of PPL values within the Colsterworth square to compare the two OS maps (Figures 3.17 and 3.18). There is far greater variation in topography in the Derwent square, which will lead to a less intensively managed environment, a fact that will also be affected by it’s location within a National Park.
Figure 3.17: Ordnance Survey 1:50,000 map of the area of the Derwent pilot square
By displaying the metadata as in Figures 3.11, 3.12 and 3.13 we are introducing a spatial element into the analysis of the metadata and also presenting it in a way that many people will find easier to understand. Such maps allow easy comparison between areas, as they are shown above and also between the metadata and the final map. The semi natural area has greater variation in parcel size and a larger parcel size on average, which needs to be considered when displaying uncertainty data. It may be important then to combine information such as counts of parcels along with the maps. This approach of offering the user multiple ways to interact with the metadata is explored in detail by Blenkinsop et al (2000).
A further point of interest is the effect of parcel size on the impression given by these maps. There are 144 parcels from the Colsterworth square shaded in the darkest category in Figure 3.11 as opposed to just 34 in the Derwent square, but those parcels also have a mean size more than three times greater than those in the Derwent square, so the visual impression of much higher values in the Colsterworth square is magnified. On the whole, mean parcel size is actually slightly higher in the Derwent square than the Colsterworth (4.71ha and 4.47ha respectively), though the area values have a far greater standard deviation (13.64 as opposed to 4.58). So the parcel sizes are more extreme and if the dark colours are in small parcels the impact on the user is altered, and the comparison between the squares is not as easy.

These diagrams do show a pitfall with visualisation techniques and that is how it is possible to display data in an inappropriate way thus affecting how the user would interpret that data (Ehlschlaeger, et al, 1997). For example, in Figure 3.11 the use of a fixed equal interval colour scheme, chosen here to allow direct comparison, means that the large difference in the numbers of parcels with no data is not conveyed at the Pix2 and Pix3 levels. This would imply that the user would need to have a certain level of knowledge or experience in using geographical data in order to make the most of such tools, and it would be desirable in that case to give the user some control over the display method.

Mapping the locations of KBCs could be useful in some circumstances, though as Section 3.4.1.2 has described, it is more restrictive than actually examining the per-parcel likelihoods and the PPL values directly. It would however be interesting to see if there were clustering of probability aggregation KBCs caused by something other than clustering of particular land cover types. Any such clustering could point to areas of problems with the imagery due to atmospheric properties. Even clusters that are related to the underlying land cover type could
indicate areas of that land cover type that are spectrally different to the training data for some reason.

3.4.4 Visualisation of support for final classification

When the locations of parcels supported by Pixl are mapped (Figure 3.15) the differences between the squares are again clear. The pattern displayed in the Derwent square is dominated by the manually filled voids in the data (Figure 3.13a), though the areas that do not support the final classification go beyond the areas of manual fill, they very much follow the overall pattern of two diagonal stripes described earlier. Areas of support follow generally the areas that have high values in Pixl (Figure 3.11a), being the south west and north east corners and the central diagonal stripe from north west to south east. The clear exception to this is the area on the middle of the western side of the square, which does not show support but does have high Pixl values. This square has 34 parcels with a Pixl value higher than 80% and 15 of those do not support the final classification. In contrast the Colsterworth square has 144 parcels where Pixl is above 80% and just one does not support the final classification.

In the Colsterworth square the parcels without supporting PPL are scattered across the square, displaying no obvious clustering. Given this, it is more likely that random processes and errors are creating the differences between the parcel and pixel classifications.

3.4.5 Simulation

The simulations created here are modelling the uncertainty within the LCM2000 by generating a number of alternative potential realisations of the land cover, all of which are...
Chapter 3: Exploration of uncertainty within LCM2000 metadata

calculated from the object-based metadata. Such techniques have been used in many studies of uncertainty (Ehlschlaeger et al., 1997; Fisher 1991b, 1993; Goodchild et al., 1992) though they are generally based on dataset level quality information, as that is all that is available to the researcher. By creating such potential realisations the researcher is able to statistically analyse the output to assess the impact of uncertainty in the data. Such realisations are particularly useful in modelling as a series of them can be used as inputs to separate runs of a model and the outputs analysed to assess the impact of uncertainty in the data on the output of the model. Animation techniques can however allow these realisations to be directly examined and provide a powerful tool to assist users in the exploration of the uncertainty in the data. This direct examination can allow the researcher to see spatial patterns or relationships in the uncertainty that had not been previously considered, thus making it a useful element in the research process as described by DiBiaise (1990).

Ehlschlaeger et al. (1997) describe how it is important when animating alternative realisations that the transition between images is an important consideration in producing a smooth animation. This is certainly the case when producing a complex 2.5D animation of sea level rise as they were. The animations created here (Appendix 3) are not as complex as they are representing a much more straightforward 2D scenario where the most important consideration is the length of time that parcels maintain the same colour and the number of changes in colour.

Watching the animation of the Colsterworth square gives a general impression of quite high uncertainty as there is considerable change between the images. A large amount of this change in the subclass level animation is changing between different types of arable category so the level of change appears somewhat less in the broad habitat level animation. The village of Colsterworth does retain an urban or suburban land cover type consistently.
Chapter 3: Exploration of uncertainty within LCM2000 metadata

throughout the animation though and it is this sort of observation that shows the worth of such techniques.

3.5 Conclusion

This chapter has looked at ways that the metadata within LCM2000 can be used, as it stands, to describe and visualise object level uncertainty. It is clear from this work that it is possible to extract useful information from the product as it is delivered, to assist users in assessing how fit LCM2000 is for their use. There are further possibilities, however, to use the metadata to produce new information, beyond that provided, both statistically and spatially (see Chapters 5 and 6). The ideas and techniques explored in this chapter could equally be extended to such new information.

In all of the analyses discussed here greater emphasis has been placed on the PPL values rather than the per-parcel likelihoods. This is partly because the PPL is a standard element of the metadata supplied as part of one of the LCM2000 products, but more importantly because the per-parcel data is not independent of the final classification, it is simply reporting further detail from the classification process. What is being reported in the per-parcel data is effectively the relationship between the parcel in question and the training data used in the classification. Training areas are chosen to be the most representative of that category (Zhang and Goodchild, 2002; p71) and will be more similar to one parcel than to another, even though both parcels would fit the criteria for that category. In this case the classifier would rightly detect differences between them, but the output from the classifier will therefore be describing the strength of the similarity to the training data, not to the category. Reporting the per-parcel data would be potentially useful to users of the map, but reporting a second independent classification is a novel approach and more the focus of this work.
Blenkinsop et al (2000) reviewed several different visualisation methods to see how well users were able to perceive the uncertainty being displayed by these methods. This was done using interactive software that produces several different types of visualisation and it was found that users with different levels of experience were generally able to successfully interpret the information on uncertainty, showing the value of such methods. It was found however that the use of grey-scale images and histograms of classification accuracy were the most successful technique. The images of PPL and parcel likelihood values shown in Figures 3.4, 3.5, 3.6, 3.7, 3.11 and 3.12 are much like those used by Blenkinsop et al. Random animations similar to those of Appendix 3 were also shown to have moderate success in aiding accurate interpretation.

Visualisation can be a useful tool for exploring data uncertainty and also communicating it to others, but even in this short discussion of the topic it is clear that while it provides many possibilities there are also many display difficulties in using such techniques effectively. A logical and desirable continuation of the animation described here is to create a more complete simulation by embedding the algorithm into the viewing software to generate a single constantly changing image based on the metadata, rather than an animation of a number of realisations. Unfortunately, time considerations have meant that it has not been possible to explore this possibility further.
Chapter 4: Cumulative evidence and supervaluation

4.1 Introduction

In the previous chapters it has been shown that the object level metadata reported within the LCM2000 does not all relate directly to existing standards for reporting data quality, and how it can be utilised to describe and visualise uncertainty in the map. This has demonstrated the usefulness of the metadata. Thus far however, only contextual information has been used to draw conclusions about the validity of the conclusions drawn from the analyses and descriptions. There is therefore a need to test that the metadata is providing further information about the relationship of the LCM2000 with the actual land cover of the UK.

In order to assess the potential effectiveness of the object-based metadata available within LCM2000, with respect to understanding attribute accuracy within the map, certain categories within the database are compared with what is termed here a cumulative evidence analysis. This analysis takes evidence that the category is present from a number of databases and calculates the number of those databases that agree on the presence of the land cover type in question, for all locations within the study area. The extent of woodland, urban areas and water are assessed using different mappings of those phenomena, and the number of times a location was identified as the named cover type is then compared with the support for the proposition of its presence in the metadata.

4.1.1 Supervaluation and cumulative evidence

The inspiration behind cumulative evidence analysis springs from the concept of supervaluation, which is a theory from the fields of philosophy and natural language (Fine, 1975; Kamp 1975). Supervaluation assumes that there is more than one acceptable interpretation of a vague object or idea and no interpretation is considered the correct one (Kulik, 2001). Such interpretations can relate to concepts or even words and a supervaluation
analysis can assist in the understanding of vagueness surrounding it. This can and has been utilised in understanding vagueness in the presence or extent of geographical objects (Bennett, 2001b; Kulik, 2001; Varzi, 2001). In this case an interpretation can be seen as a mapping of a geographical object, thus defining the location and extent of that object. Although the following description relates to supervaluation in general it will use specific terms and examples relating to its use with respect to geographical objects.

Each interpretation in a supervaluation analysis is an attempt to create a precise definition of the object in question, and so is termed a precisification. All of the admissible precisifications are allocated a truth-value, such that all areas included in the object according to that precisification are deemed true, and bringing these values together is called a supervaluation (Kulik, 2001). The supervaluation allocates a truth-value to an object, which is obtained from a function of the truth-values according to the individual precisifications (Varzi, 2001). In all the locations where an object has a positive truth-value in all the precisifications, these locations are termed supertrue (or alternatively termed unequivocally true; Bennett, 2001b). The locations where the converse situation occurs, where all precisifications have a negative truth-value, are termed superfalse. All remaining locations must therefore have a positive truth-value in some precisifications and a negative truth-value in others. In this scenario classical supervaluation theory assigns no truth-value at all.

In assessing the cumulative evidence the number of truth-values from the precisifications (or number of individual pieces of evidence in the form of spatial databases) are summed at all locations. Where the number of precisifications used in an analysis is \( n \), the value assigned by the cumulative evidence analysis will be between zero and \( n \). A count of evidence is the output of the analysis, giving a gradation of values between \( n \) in those locations where all
Chapter 4: Cumulative evidence and supervaluation

evidence agrees on the presence of the feature in question and zero where all evidence agrees on the absence of the feature.

Some care evidently must be taken over what precisifications are deemed to be admissible to the valuation as all evidence, or each precisification, has equal value. Varzi (2001) suggests that constraints must be applied in order to control the validity of precisifications. When analysing presence of geographical objects in experimental situations it is possible to control the production of the precisifications, but when you are studying a large area, such as the UK, it is not possible to create a series of new independent datasets in such a controlled way. In this situation you require extant datasets to form the basis of the analysis, meaning that such control is impossible and it is inevitable that some datasets are more appropriate than others for the particular analysis. Given this situation it is important to know as much as possible about the different datasets used in order to understand how they might affect the output.

The output of the cumulative evidence analysis becomes an ideal dataset against which to compare the metadata within LCM2000 as it gives a measure of how likely it is that a specific land cover type occurs at a specific location. LCM2000 metadata gives an alternative measure of the likely presence of a land cover type at that location and so the two can be directly compared.

An example application of the cumulative evidence approach is outlined below, showing that it not only provides spatially explicit estimates of the extent of a land cover, but that it can also be combined with further data relating to land management regimes in order to assess the impacts of that management. The example relates to estimating the extent of woodland in the National Forest (see Section 1.3) and assessing the impact that planting initiatives are having on this. In doing this, the approach still allows the portrayal of uncertainty within the
estimates by presenting a range of estimates of woodland extent from very conservative to more liberal.

The different databases used in the cumulative evidence analysis are outlined in the following section, followed by a detailed description of the methodology. Results are then presented, first from the cumulative evidence analysis itself, then from the comparison of these with the metadata and finally for the application of the analysis to assess the extent of woodland within the National Forest. The results show that the cumulative evidence analysis can give spatially explicit estimates of the extent of a land cover type and importantly that there is a strong relationship, at least in the two cases for which good data was obtainable, between the output of the analysis and the LCM2000 metadata. Finally the analysis output is shown to be of practical use in guiding management policy with relation to land cover and given the strong relationship with the metadata, it is possible to see how the metadata alone could be used in a very similar way.

4.2 **Input databases for cumulative evidence analysis**

Table 4.1 outlines the datasets used in the analysis of attribute uncertainty, and clearly shows the variety of different intended scales and purposes in their creation. As they are to be used as evidence of the presence of specific land cover types, it would be preferable to use datasets that have a shared ontology, but it is unlikely that many mappings of the *same* area at the *same* scale for the *same* purpose at the *same* time (a fully shared ontology) have ever occurred in more than a very limited number of controlled experimental situations (Edwards and Lowell, 1996; Middelkoop, 1990). It certainly has not occurred for the land cover of Great Britain.
Chapter 4: Cumulative evidence and supervaluation

The ideal situation of a fully shared ontology between the inputs is therefore not possible, but the aim of the work reported here was to utilise as many extant datasets as possible that report aspects of land cover over a large area. The premise being that the more datasets that are in agreement about the presence of a particular phenomenon, the lower will be the uncertainty of that phenomenon occurring at that location. We refer to this as a cumulative evidence analysis. As noted above the datasets available are heterogeneous in their ontology, which will obviously lead to differences, but all purport to represent the extent of the phenomenon in question within the study area. In these aspects the ontology is shared. Most were also created or last updated over a similar timescale, with the National Inventory of Woodland and Trees (NIWT) and the Ancient Woodland Inventory (AWI) both completed in 2000, Ordnance Survey (OS) Meridian being updated in 2001 and Bartholomew’s dataset undergoing a major enhancement in 1998. The Urban Settlement project was started in 2001 and LCM2000 is created from satellite imagery dating from the late 1990’s. The final two datasets are slightly older with the census wards being those of the 1991 census and the Land Cover Map of Great Britain (LCMGB) created from satellite imagery from 1988 to 1990. There is therefore an issue with currency of data, which cannot be avoided, leading to the possibility that land cover can change between production of the different datasets. A majority of such change is likely to be over small areas and so would be detected by those datasets with a fine resolution, but is less likely to affect the coarse resolution products. Such a change would lead to some of the datasets designating an area as the target land cover class and some would not, so there would be some evidence for that phenomenon at that location, but it would not be certain as not all datasets are in agreement. This is still an indication of uncertainty from the analysis, which is the aim given the lack of a definitive, totally accurate map.
Resolution is another property of the datasets that is not shared (Table 4.1), nor is the purpose for creating the datasets, with some products aiming to create an exhaustive map of land cover (LCMGB, LCM2000), some being general topographic datasets (Meridian, Bartholomew’s), one being based on administrative boundaries (census wards) and others being very specific analyses of the extent of particular phenomena (AWI, NIWT, Urban Settlements).

As far as possible it has been ensured that the datasets are not directly derived from one another. Bartholomew’s data was originally sourced from OS 1:250,000 mapping, but has since been updated from photographic and survey evidence by the publishers (Gill, personal communication, 2003). AWI is also sourced from OS mapping, but at 1:25,000 scale and using 1920’s maps as the base for its production. The Urban Settlement boundaries were sourced from current OS 1:10,000 maps with other datasets being produced independently from aerial photography and satellite imagery. The extent of the forest area designated as woodland, urban and water are given in Table 4.2, showing how the differences between the datasets detailed above have affected the mapping of these three features over the same area.
Table 4.1: The sources and characteristics of datasets used as inputs to the cumulative evidence analysis.

<table>
<thead>
<tr>
<th>Dataset title</th>
<th>Copyright owner / originating organisation</th>
<th>Resolution/Associated Scale</th>
<th>Availability</th>
<th>Data format</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bartholomew's dataset</td>
<td>Collins Bartholomew</td>
<td>1:200,000</td>
<td>Licence agreement</td>
<td>Vector</td>
<td>Enhanced and updated version of 1:250,000 OS data.</td>
</tr>
<tr>
<td>Meridian dataset</td>
<td>Ordnance Survey</td>
<td>1:250,000</td>
<td>Licence agreement</td>
<td>Vector</td>
<td>Originated from OS Strategi, a 1:250,000 product.</td>
</tr>
<tr>
<td>Land Cover Map of Great Britain (1990; LCMGB)</td>
<td>Centre for Ecology and Hydrology</td>
<td>1250 m² Minimum mapping unit (MMU)</td>
<td>Licence agreement</td>
<td>Raster</td>
<td>25m raster grid created from satellite imagery using maximum likelihood classifier</td>
</tr>
<tr>
<td>Land Cover Map 2000 (LCM2000)</td>
<td>Centre for Ecology and Hydrology</td>
<td>1250 m² (MMU)</td>
<td>Licence agreement</td>
<td>Vector</td>
<td>Parcel-based dataset created from segmented satellite imagery</td>
</tr>
<tr>
<td>National Inventory of Woodland and Trees (NIWT)</td>
<td>Forestry Commission</td>
<td>1:25,000 – 1:50,000 maps; 2 ha MMU</td>
<td>Licence agreement</td>
<td>Vector</td>
<td>1:25,000 aerial photographs and ground survey</td>
</tr>
<tr>
<td>Ancient Woodland Inventory (AWI)</td>
<td>English Nature</td>
<td>1:25,000 maps; 2 ha MMU</td>
<td>Public Domain</td>
<td>Vector</td>
<td>Various paper maps, with base maps from the 1920's</td>
</tr>
<tr>
<td>Urban Settlement boundaries</td>
<td>Office for National Statistics</td>
<td>1:10,000</td>
<td>Public Domain</td>
<td>Vector</td>
<td>Sourced from OS 1:10,000 mapping</td>
</tr>
<tr>
<td>Census Urban Wards</td>
<td>EDINA</td>
<td>1:2,500 – 1:50,000 maps</td>
<td>Licence agreement</td>
<td>Vector</td>
<td>Based on census wards defined as urban or rural</td>
</tr>
</tbody>
</table>
4.3 Method

All datasets for each category were converted into binary presence-absence maps on a 5m grid and added together to calculate the number of times the phenomenon in question was present. The output was a map that gave a value to every location in the study area between zero (all datasets agreed that the category is absent) and the total number of datasets (all agreed that the category is present), which is four in the case of open water and six for woodland and urban areas. This map of values was taken to be a representation of the level of certainty in the presence of that category. Middelkoop (1990) and Edwards and Lowell (1996) utilised a similar method, of combining different realisations or interpretations of the same set of objects, thus building a map of the level of agreement within those realisations. In both cases they were analysing boundary uncertainty using multiple manual interpretations of images. Molenaar (1998) also used such a method and was able to calculate a fuzzy set membership value for all areas, thus indicating the reliability of the classification. Here the
ranked value from the overlay operation was taken as a relative measure of certainty of the mapped cover type (Table 4.3).

These maps were then compared with both the per-parcel and the per-pixel classification metadata from LCM2000, in order to see if patterns of uncertainty were similar between the two analyses. A list was generated of all the LCM2000 parcels that overlapped with areas of each cumulative evidence value. In the comparison any parcels that had been subjected to KBCs were removed from the per-parcel analysis, as the corrections would render relevance of the spectral probability value questionable. The parcels subject to KBCs were not removed from the per-pixel analysis as this was derived from a separate classification procedure and so was not affected by the KBCs. Parcels that had been filled manually due to such problems as cloud cover in the satellite images were removed however as these contained no per-pixel data. The number \( n \) of parcels overlapping with each value was counted and the per-pixel and per-parcel metadata within those parcels was analysed to assess if the metadata is giving direct information regarding the attribute uncertainty of LCM2000.

The results of the cumulative evidence analysis are also utilised to assess the extent of woodland within the National Forest. Using this technique allows the woodland extent to be analysed spatially as the method can be used across a wide variety of scales. The National Forest Company has divided the forest area into six regions. These are areas that have been created for management reasons and basically represent geographically different regions within the forest representing differences based on historical land use and character, though they were not created with any rigorous method. As they are designed to be management units it is useful to analyse them individually to allow comparison and assess the affects of management within each region. The analysis is broken down further still to the parish level as it could provide useful information to Parish councils or allow the National Forest...
Chapter 4: Cumulative evidence and supervaluation

Company to break down their management units and assess results within those units. A map of the parishes within the National Forest was compiled from UKBORDERS data. It was necessary to discard some small areas around the boundary of the forest, as these were small sections of parishes, probably created by differences between boundary datasets, and so too small to be useful within the analysis. A map of the 70 parishes used in the analysis is given in Figure 4.14. In the UKBORDERS data the parishes of Burton upon Trent, Stanton and Newhall, Swadlincote and Church Gresley were amalgamated into a single polygon. This is shown as Area A in Figure 4.14, as it was not possible to differentiate the separate parishes. The cumulative evidence output was split across the parishes, as described above for the whole forest and for the landscape regions.

As well as breaking down the cumulative evidence results spatially to assist management, they can be compared with other information regarding management history and planting schemes within the forest to derive estimates of woodland cover and to assess the future impact of current management practices. The National Forest Company form agreements with landowners to maintain existing woodland or to undertake new planting schemes, with grants and subsidies often available, such as from the Woodland Grant Scheme run by the Forestry Commission. The locations of these areas were provided by the National Forest Company for this work and are called the Woodland Regions dataset. Many of the areas in this dataset have been planted over recent years, so are not likely to be sufficiently developed as woodland to be detected and classified as such from satellite imagery. Analysing likely woodland extents from the cumulative evidence analysis along with this management data will therefore allow us to assess the likely impact of that management. This was achieved by assessing the values from the cumulative evidence within the Woodland Regions and those that are unlikely to be woodland we can assume will become woodland as long as the...
management regime is not altered. From this, estimates of the current extent of woodland can be compared with estimates for the future extent based on current management practices.

4.4 Results of comparison of cumulative evidence analysis with metadata

4.4.1 Metadata from per-pixel classification - woodland

The outcome of the assessment of uncertainty regarding the presence of woodland (Figure 4.1) showed that total agreement in the presence of woodland covered just 0.83% of the study area (Table 4.3). This is considerably less than the lowest area given by any of the individual data sets, which was 2.61% of the study area (Table 4.1). The area allocated value 4 and above, meaning that more than half of the data sets agreed that woodland was present, still only accounts for 2.57%, illustrating a considerable level of disagreement between the data sets.

Table 4.3: Output of cumulative evidence analysis for woodland, showing the percentage of the National Forest area allocated to each value.

<table>
<thead>
<tr>
<th>Ranked value</th>
<th>Proportion of National Forest in woodland analysis</th>
<th>Proportion of National Forest in urban areas analysis</th>
<th>Proportion of National Forest in open water analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0.83%</td>
<td>2.54%</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.79%</td>
<td>2.81%</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.95%</td>
<td>2.08%</td>
<td>0.17%</td>
</tr>
<tr>
<td>3</td>
<td>1.84%</td>
<td>3.06%</td>
<td>0.23%</td>
</tr>
<tr>
<td>2</td>
<td>2.54%</td>
<td>6.72%</td>
<td>0.29%</td>
</tr>
<tr>
<td>1</td>
<td>8.64%</td>
<td>22.36%</td>
<td>1.11%</td>
</tr>
<tr>
<td>0</td>
<td>84.44%</td>
<td>60.44%</td>
<td>98.20%</td>
</tr>
</tbody>
</table>

The per-pixel data was reviewed to give a percentage of each parcel that had been designated as a woodland category. This percentage was compared with the values generated by the cumulative evidence analysis. Parcels with each value (0–6) were selected, and the attribute
data analysed to assess the percentage of the parcel designated as woodland by the per-pixel classification. This analysis therefore reflects parcels created by the intersection process and so the number of parcels is not the same as in the LCM2000, and the same parcel (and so the same metadata) can appear in more than one value. Data for each of the values is summarised in Figure 4.2.

Figure 4.1: Map of the National Forest showing total agreement on the presence of woodland or value 6 (black), total agreement on the absence of woodland, or value 0 (white) and non-agreement, or values 1-5 (grey) from the cumulative evidence analysis.
None of the per-pixel datasets were normally distributed (verified using an Anderson-Darling test), so a nonparametric one-sample sign test was used to calculate the 95% confidence intervals (CIs), based on median values. This shows that values 0 and 1 are indistinguishable and that there is a slight overlap in confidence intervals between values 3 and 4, and values 4 and 5. Confidence intervals exhibit no overlap at the 85% level, other than between the two lowest values, therefore per-pixel values could predict the ranked value of a location with this level of confidence.

4.4.2 Metadata from per-pixel classification - urban areas

Urban areas were analysed using the same process as described above. The proportion of the study area designated the different values is given in Table 4.3, and the map of the output is shown in Figure 4.3. Total agreement in the presence of urban areas (2.54%; value 6, Table
4.3) is again considerably less than the lowest area given by any individual dataset, which is 7.76% (Table 4.3). The area allocated 4 or above, accounts for 7.43% of the study area.

Figure 4.4 shows that there is again a clear relationship between the cumulative evidence values and the per-pixel percentages. Despite the range of percentages being zero to 100 for each value, the 95% CIs display no overlap other than between values zero and one (Figure 4.4). This is also true for CIs at the 99% level.

**Figure 4.3:** Map of the National Forest showing total agreement on the presence of an urban category, value 6 (black); total agreement on the absence of urban, value 0 (white); non-agreement, values 1-5 (grey).
Figure 4.4: Boxplot showing per-pixel percentage values for the urban category from parcels within areas of different values generated by the cumulative evidence analysis (box = inter-quartile range, vertical line = median value, o = mean value, ■ ■ ■ = 95% confidence interval, n = number of land parcels intersecting with areas of that value).

4.4.3 Metadata from per-pixel classification - open water

The results from the analysis of open water (Table 4.3 and Figures 4.5 and 4.6) are much less satisfying than those from woodland and urban areas. This is because there is not very much open water within the study area (Table 4.3) leading to very small sample sizes for locations allocated values of three or four in the cumulative evidence analysis. For this reason the results cannot be seen as in any way conclusive, but they do still exhibit a pattern of increasing per-pixel percentages in relation to higher values from increasing cumulative evidence.
Only 0.17% of the study area has total agreement in the presence of open water (Table 4.3). The proportion of the study area allocated a value of three or four is 0.40%, which is slightly less than the smallest proportion designated as water by any of the individual datasets (Table 4.2).

Again the two lowest values are indistinguishable from one another and there is some overlap between the 95% confidence intervals (Figure 4.6). There is no overlap of confidence intervals at the 85% level other than the two lowest values, which still have a range of zero at this level.
Figure 4.6: Box plot showing per-pixel percentage values for the open water category from parcels within areas of different values generated by the cumulative evidence analysis (box = inter-quartile range, vertical line = median value, $\circ$ = mean value, $\ldots$ = 95% confidence interval, $n$ = number of land parcels intersecting with areas of that value).

4.4.4 Metadata from per-parcel classification – woodland, urban and open water

The spectral probability values were split into those that had been finally categorised as the target land cover class in the LCM2000, and those that had not. Figures 4.7, 4.8 and 4.9 show the output of the comparison between the spectral probability percentages and the cumulative evidence values for each of the three target land cover types. None of these results show a clear relationship.
Chapter 4: Cumulative evidence and supervaluation

**Figure 4.7:** Per parcel spectral probability percentages for parcels categorised as woodland (a) and non-woodland (b) within areas of different values from the woodland cumulative evidence analysis (box = inter-quartile range, vertical line = median value, o = mean value).

**Figure 4.8:** Per parcel spectral probability percentages for parcels categorised as urban (a) and non-urban (b) within areas of different values from the urban cumulative evidence analysis (box = inter-quartile range, vertical line = median value, o = mean value).

**Figure 4.9:** Per parcel spectral probability percentages for parcels categorised as water (a) and non-water (b) within areas of different values from the water cumulative evidence analysis (box = inter-quartile range, vertical line = median value, o = mean value).
4.5 **Results from the analysis of the extent of woodland in the National Forest**

4.5.1 **The National Forest**

Cumulative evidence analysis can show us likely areas of woodland in the National Forest (Figure 4.1; Table 4.3). This can be broken down into smaller spatial units, such as the Landscape Regions used by the National Forest Company as management units (Figure 4.10; Section 4.5.2). The results can also be used in conjunction with other data regarding management regimes within the area of interest.

The values from the cumulative evidence analysis were compared with the Woodland Regions dataset supplied by the National Forest Company (Table 4.4; Figure 4.13). Over 90% of these areas have a value of zero or one (Table 4.4), meaning that they are in areas where there is little or no evidence for the presence of woodland prior to the new management schemes being put into place. Therefore it shows that almost all of these areas represent an increase in woodland cover, or at least will do so if the management continues unchanged for a period of time.

<table>
<thead>
<tr>
<th>Ranked Value</th>
<th>Total Area (km²)</th>
<th>Percentage of total Woodland Region area</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0.016</td>
<td>0.04%</td>
</tr>
<tr>
<td>5</td>
<td>0.056</td>
<td>0.15%</td>
</tr>
<tr>
<td>4</td>
<td>0.323</td>
<td>0.84%</td>
</tr>
<tr>
<td>3</td>
<td>1.047</td>
<td>2.74%</td>
</tr>
<tr>
<td>2</td>
<td>1.793</td>
<td>4.69%</td>
</tr>
<tr>
<td>1</td>
<td>13.416</td>
<td>35.08%</td>
</tr>
<tr>
<td>0</td>
<td>21.596</td>
<td>56.46%</td>
</tr>
</tbody>
</table>

The cumulative evidence analysis suggests that the level of woodland cover in the National Forest is anywhere between 0.83% and 15.59% (values 1 – 6, Table 4.3). However, it is probably more reasonable to view the figure to be between 0.83% and 6.95% (values 2 – 6,
Table 4.3), as problems such as misregistration of datasets mean that when there is only a single piece of evidence that woodland is present (value 1), there is considerable doubt over the validity of that evidence (see Section 4.6.1.1). If it is taken that areas with a value of two or greater are potentially woodland then the areas identified in the Woodland Regions dataset increases the possible range by 35km$^2$ (Woodland Region areas with value zero or one; Table 4.4). That means a range of between 7.77% and 13.89% of the National Forest is covered by woodland or recently planted trees. This estimate can be compared with the proportion of England that is wooded in 2001 which was 8.4% (Countryside Agency, 2002), showing that the management regime is likely to ensure that the National Forest is considerably more wooded than the national average.

### 4.5.2 Landscape Regions

The cumulative evidence output was then split across the landscape regions identified by the National Forest Company (shown in figure 4.10). Table 4.5 and Figure 4.11 show that there is a considerable difference between the landscape regions with respect to woodland cover (see also Figure 4.10). Charnwood and Calke Uplands have 17.34% and 15.99% respectively allocated values of two or above, whereas Midlands Coalfield and Trent Valley have only 1.50% and 2.10% respectively (see Table 4.8 for detailed figures).

<table>
<thead>
<tr>
<th>Landscape Region</th>
<th>Value 6 (%)</th>
<th>Value 5 (%)</th>
<th>Value 4 (%)</th>
<th>Value 3 (%)</th>
<th>Value 2 (%)</th>
<th>Value 1 (%)</th>
<th>Value 0 (%)</th>
<th>Values 2-6 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calke Uplands</td>
<td>2.68</td>
<td>2.63</td>
<td>1.81</td>
<td>3.41</td>
<td>5.45</td>
<td>10.49</td>
<td>73.52</td>
<td>15.99</td>
</tr>
<tr>
<td>Charnwood</td>
<td>1.95</td>
<td>2.18</td>
<td>3.02</td>
<td>4.82</td>
<td>5.37</td>
<td>12.29</td>
<td>70.36</td>
<td>17.34</td>
</tr>
<tr>
<td>Mease Lowlands</td>
<td>0.43</td>
<td>0.25</td>
<td>0.45</td>
<td>0.94</td>
<td>1.03</td>
<td>8.18</td>
<td>88.72</td>
<td>3.11</td>
</tr>
<tr>
<td>Midlands Coalfield</td>
<td>0.00</td>
<td>0.00</td>
<td>0.08</td>
<td>0.49</td>
<td>0.93</td>
<td>7.57</td>
<td>90.92</td>
<td>1.50</td>
</tr>
<tr>
<td>Needwood</td>
<td>1.14</td>
<td>0.84</td>
<td>1.20</td>
<td>2.31</td>
<td>3.19</td>
<td>6.34</td>
<td>84.98</td>
<td>8.68</td>
</tr>
<tr>
<td>Trent Valley</td>
<td>0.00</td>
<td>0.00</td>
<td>0.03</td>
<td>0.58</td>
<td>1.49</td>
<td>8.33</td>
<td>89.57</td>
<td>2.10</td>
</tr>
</tbody>
</table>
Figure 4.10: Cumulative evidence values from the woodland analysis within Landscape Regions (red boundaries).

Figure 4.11: Percentage of Landscape Regions allocated to each value from cumulative evidence analysis – value 0 is not shown.
The total area figures for the landscape regions are slightly different to those for the whole National Forest. This is, at least in part, due to the differences in boundary datasets of the entire forest area and that of the Landscape Regions. Although these differences are only slight all figures are presented here as proportions for ease of comparison.

The woodland regions dataset was compared with the separate landscape region data in the same way as described above for the whole forest. Charnwood has much higher percentages of the planted areas (from Woodland Regions dataset) located in areas of values two to six, than the other landscape regions (Table 4.6). This is probably an indication that management schemes are taking place in areas associated with existing woodland. Charnwood does have nearly double the amount of woodland of any of the other landscape regions (14.49 km² given value 2-6; Table 4.8).

Estimations of the possible range of woodland cover in each landscape region were made in the same way as was done above for the whole forest. The results are shown in Table 4.7 and Figure 4.12, and show clearly that the planting schemes are having a large impact on the potential range of woodland in Mease Lowlands, Midlands Coalfield and Trent Valley; also that they have had the least impact on Charnwood and Needwood. The impact of management can be seen in Figure 4.12, by looking at the difference between the two lines for each region. This is also shown in Figure 4.13, which depicts all areas with a value between two and six from the original analysis with all recently planted areas within areas of value zero or one. Figure 4.12 also gives a clear indication of how variable the evidence is. Short bars on the graph mean that there is little difference in the area assigned a six and the area assigned a value of two or more. The regions with larger areas of woodland also have quite wide potential ranges in woodland extent, which is likely to be a function of the differences in the datasets leading to disagreement between them.
### Table 4.6: Comparison of Woodland Region dataset with ranked values from cumulative evidence analysis, split by landscape region

<table>
<thead>
<tr>
<th>Ranked Value</th>
<th>Total Area (km²)</th>
<th>Percentage of total Woodland Region area</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total Area (km²)</td>
<td>Percentage of total Woodland Region area</td>
</tr>
<tr>
<td></td>
<td>Total Area (km²)</td>
<td>Percentage of total Woodland Region area</td>
</tr>
<tr>
<td></td>
<td>Total Area (km²)</td>
<td>Percentage of total Woodland Region area</td>
</tr>
</tbody>
</table>

### Table 4.7: Potential range in woodland cover within each landscape region and the effect on that range once woodland region areas valued zero or one are added

<table>
<thead>
<tr>
<th>Landscape Region</th>
<th>Potential range of woodland cover (not including planting schemes)</th>
<th>Potential range of woodland cover (including planting schemes)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower limit (value 6)</td>
<td>Upper limit (values 2-6)</td>
</tr>
<tr>
<td></td>
<td>Lower limit (value 6 + woodland region)</td>
<td>Upper limit (values 2-6 + woodland region)</td>
</tr>
</tbody>
</table>

114
Figure 4.12: Potential range in woodland cover within each landscape region. Upper bar for each landscape region shows potential range without planting schemes; Lower bars show potential range including planning schemes within areas of value 0 and 1.
Figure 4.13: Landscape regions (red boundaries) showing areas allocated a value between 2 and 6 (black) with recently planted areas within areas allocated a value of 0 or 1 (grey)
### Table 4.8: Output of cumulative evidence analysis showing the percentage of each Landscape Region allocated to each value.

<table>
<thead>
<tr>
<th>Ranked value</th>
<th>Needwood</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Area km²</td>
<td>Proportion of study area</td>
<td>Cumulative proportion of study area</td>
<td>Proportion of study area</td>
</tr>
<tr>
<td>6</td>
<td>0.80</td>
<td>1.14%</td>
<td>1.14%</td>
<td>1.14%</td>
</tr>
<tr>
<td>5</td>
<td>0.59</td>
<td>0.84%</td>
<td>1.98%</td>
<td>1.98%</td>
</tr>
<tr>
<td>4</td>
<td>0.84</td>
<td>1.20%</td>
<td>3.18%</td>
<td>3.18%</td>
</tr>
<tr>
<td>3</td>
<td>1.62</td>
<td>2.31%</td>
<td>5.49%</td>
<td>13.88%</td>
</tr>
<tr>
<td>2</td>
<td>2.23</td>
<td>3.19%</td>
<td>8.68%</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4.44</td>
<td>6.34%</td>
<td>15.02%</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>59.54</td>
<td>84.98%</td>
<td>100.00%</td>
<td>84.98%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Calke Uplands</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>1.43</td>
<td>2.68%</td>
<td>2.68%</td>
</tr>
<tr>
<td>5</td>
<td>1.40</td>
<td>2.63%</td>
<td>5.31%</td>
</tr>
<tr>
<td>4</td>
<td>0.96</td>
<td>1.81%</td>
<td>7.12%</td>
</tr>
<tr>
<td>3</td>
<td>1.82</td>
<td>3.41%</td>
<td>10.54%</td>
</tr>
<tr>
<td>2</td>
<td>2.90</td>
<td>5.45%</td>
<td>15.99%</td>
</tr>
<tr>
<td>1</td>
<td>5.58</td>
<td>10.49%</td>
<td>26.48%</td>
</tr>
<tr>
<td>0</td>
<td>39.11</td>
<td>73.52%</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chamwood</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>1.63</td>
<td>1.96%</td>
<td>1.96%</td>
</tr>
<tr>
<td>5</td>
<td>1.82</td>
<td>2.18%</td>
<td>4.13%</td>
</tr>
<tr>
<td>4</td>
<td>2.52</td>
<td>3.02%</td>
<td>7.15%</td>
</tr>
<tr>
<td>3</td>
<td>4.03</td>
<td>4.82%</td>
<td>11.97%</td>
</tr>
<tr>
<td>2</td>
<td>4.49</td>
<td>5.37%</td>
<td>17.34%</td>
</tr>
<tr>
<td>1</td>
<td>10.28</td>
<td>12.30%</td>
<td>29.64%</td>
</tr>
<tr>
<td>0</td>
<td>58.81</td>
<td>70.36%</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mease Lowlands</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0.25</td>
<td>0.43%</td>
<td>0.43%</td>
</tr>
<tr>
<td>5</td>
<td>0.15</td>
<td>0.25%</td>
<td>0.68%</td>
</tr>
<tr>
<td>4</td>
<td>0.27</td>
<td>0.46%</td>
<td>1.14%</td>
</tr>
<tr>
<td>3</td>
<td>0.55</td>
<td>0.94%</td>
<td>2.08%</td>
</tr>
<tr>
<td>2</td>
<td>0.61</td>
<td>1.03%</td>
<td>3.11%</td>
</tr>
<tr>
<td>1</td>
<td>4.81</td>
<td>8.18%</td>
<td>11.28%</td>
</tr>
<tr>
<td>0</td>
<td>52.21</td>
<td>88.72%</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Midlands Coalfield</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0.00</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>5</td>
<td>0.00</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>4</td>
<td>0.15</td>
<td>0.08%</td>
<td>0.08%</td>
</tr>
<tr>
<td>3</td>
<td>0.92</td>
<td>0.49%</td>
<td>0.57%</td>
</tr>
<tr>
<td>2</td>
<td>1.76</td>
<td>0.93%</td>
<td>1.50%</td>
</tr>
<tr>
<td>1</td>
<td>14.26</td>
<td>7.57%</td>
<td>9.08%</td>
</tr>
<tr>
<td>0</td>
<td>171.23</td>
<td>90.93%</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trent Valley</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0.00</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>5</td>
<td>0.00</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>4</td>
<td>0.02</td>
<td>0.03%</td>
<td>0.03%</td>
</tr>
<tr>
<td>3</td>
<td>0.28</td>
<td>0.58%</td>
<td>0.61%</td>
</tr>
<tr>
<td>2</td>
<td>0.73</td>
<td>1.49%</td>
<td>2.10%</td>
</tr>
<tr>
<td>1</td>
<td>4.08</td>
<td>8.33%</td>
<td>10.43%</td>
</tr>
<tr>
<td>0</td>
<td>43.90</td>
<td>89.57%</td>
<td>100.00%</td>
</tr>
</tbody>
</table>
4.5.3 Parishes

The output from the parish analysis is very detailed because of the large number of parishes; it is presented in Figure 4.15 and Table 4.9. Although the boundaries of the parishes and landscape regions are not necessarily related, the parishes have been roughly grouped in the graph of Figure 4.15 to generally reflect the landscape regions in which they are located. By breaking the results down into smaller geographical units it is easier to see which areas within a region have lower levels of woodland cover. If it were considered desirable to have increased woodland across the whole area then effort can be concentrated in these areas.

Figure 4.14: Parishes within the National Forest
Chapter 4: Cumulative evidence and supervaluation
Table 4.9: Ranked values from cumulative evidence anal
sbbhb

M
B
H
B
int1fBm

Alrewas
Anslow
Area A
Ashby Woulds
Ashby-de-la-Zouch
Bagworth
Bardon
Barton-Under-Needwood
Belton
Branston
Bretby
Calke
Castle Gresley
Catton
Cauldwell
Charley
Coalville
Coleorton
Coton in the Elms
Desford
Drakelow
Draycott In The C lay
Dunstall
Edingale
Foremark
Groby
Hanbury
Hartshorne
Heather
Hoar Cross
Ibstock
Linton
Loughborough
Lullington
Markfield
Measham
Melbourne
Nailstone
Netherseal
Newborough
Newtown Linford
Normanton Le Heath
Oakthorpe and Donisthorpe
Osgathorpe
Outwoods
Overseal
Packington
Ratby
Ravenstone W ith Snibstone
Repton
Rosliston
Shackerstone
Shepshed
Smisby
Snarestone
Stanton by Bridge
Stanton-Under-Bardon
Staunton Harold
Stretton
Swannington
Swepstone
Tatenhill
Ticknall
Ulverscroft
Walton Upon Trent
Woodhouse
Woodville
Worthington
Wychnor
Yoxall

0.00
0.00
0.00
0.00
0.00
0.00
2.46
0.00
0.00
0.00
2.2 0
0.00
0.00
0.00
0.00
0.00
0.00
1.61
0.00
0.00
0.00
38.99
0.00
0.00
0.00
3 .23
0.68
0.00
0 .00
8.45
0.00
0.00
2.91
0.00
0.00
0.00
0.00
0.00
2.45
0.00
3.56
0.00
0.00
0.00
0.00
0.00
0.00
8.98
0.00
14.20
0.00
0.00
0.00
2.26
0.00
21 .44
0.00
2.27
0.00
0.00
0.00
0.00
0.38
1.15
0.00
0.00
0.00
0.00
0.00
1.82

p
gfl
mSMM
B
i n li

0 .0 0
0 .0 0
0 .0 0
0 .0 0
0 .0 0
0 .0 0
2 .0 0
0 .0 0
0 .0 0
0 .0 0
1.39
0 .0 0
0 .0 0
0 .0 0
0 .0 0
0 .7 3
0.01
1.44
0.0 0
0.0 0
0.0 0
18.34
0 .0 0
0 .0 0
0 .0 0
4 .4 2
1.45
0.0 0
0.0 0
2.91
0.0 0
0.0 0
6 .3 9
0.0 0
0.0 0
0.0 0
0.0 0
0.0 0
1.39
0.0 0
4 .4 6
0.0 0
0.0 0
0.0 0
0.0 0
0.0 0
0.0 0
3.91
0 .0 0
5.5 0
0 .0 8
0 .0 0
0 .0 0
6.6 2
0.0 0
5.3 6
0.0 0
5.7 6
0.0 0
0.0 0
0.0 0
0.0 0
1.96
2.8 3
0.0 0
0.0 0
0.0 0
0.0 0
0.00
1.86

0 .0 0
0 .8 9
0 .0 9
0 .0 0
0.21
0 .4 0
3.5 2
0 .0 6
9 .7 2
0 .0 5
1.14
0.01
0.0 0
1.17
0.0 0
2 .0 9
0.5 4
1.29
0.0 0
0.0 0
0.6 3
4 .8 0
0.0 0
0.0 0
0.0 0
5.8 7
1.61
0.21
0.0 0
2.5 9
0.00
0.00
6.5 7
0.00
0.00
0.00
0.12
0.11
1.14
0.00
4.94
0.28
0.0 0
0.0 0
0.00
0.0 0
0.0 0
2.6 0
0.01
2 .1 3
0.1 3
0.0 0
0.0 3
2 .9 7
0.0 0
2.4 4
0.0 0
4 .4 3
0.0 0
0.0 0
0.00
1.16
1.98
4.7 2
0.62
0.00
0.0 0
0.0 0
1.15
1.78

119

0.0 0
2.41
0.2 9
0.3 7
0.9 3
0.6 4
5.8 4
0.3 7
9.1 7
0.1 3
3.69
1.27
2.9 3
3.06
0.01
2.96
1.14
3.33
0.00
0.11
3.28
3.44
2 .3 3
0 .18
2.19
9.20
1.35
1.57
0 .00
4.77
0.00
0 .05
8.09
0.57
0.12
0.00
1.32
0.84
1.05
1.06
6.69
0.24
0.82
0.00
0.01
1.48
0.00
2.93
2.34
2.21
0.54
0.00
1.18
3.13
0.00
1.19
0.19
6.42
0.00
0.00
0.00
3.03
3.81
7.58
1.20
8.82
0.36
0.95
2.8 3
1.77

0 .6 9
2.51
0 .5 6
1.10
2.01
0.6 3
6 .4 9
0.8 8
5 .5 9
0.9 2
6 .4 9
4 .6 4
2 .1 7
3.6 2
0.1 0
4 .8 5
1.74
8.0 2
0.0 0
0.6 9
4 .1 6
4 .7 3
2.9 2
0.6 9
3.11
8 .0 6
1.66
2.4 2
0.4 5
2 .6 4
0.3 0
0.7 0
8.9 8
0.84
0.5 7
0.12
1.71
2.1 2
1.44
2.8 8
6.1 7
0.1 5
1.13
0.0 0
0.81
0.6 3
0 .3 3
3.8 9
2 .3 0
5.9 3
1.46
0.0 0
2.0 7
2 .6 4
0.0 0
0.8 4
0.41
7.1 3
0.0 4
0.41
0.01
5.1 6
4 .8 6
8.41
1.65
7.3 2
1.21
6 .3 3
4 .1 0
2 .9 4

10.31
3.6 5
4 .4 9
12.13
8.04
11.46
16.93
8.42
8.81
9.64
8.15
13.00
3.4 4
8.7 6
3.3 0
11.23
8.4 6
9.28
15.47
10.21
6.3 4
9.84
7.49
6.41
13.40
14.50
5.73
6.65
10.48
6.9 4
10.61
18.00
18.10
2.1 8
10.77
4.02
4.3 8
7.95
4.42
3.82
12.10
3.48
9.30
0.00
2 .3 3
7.2 6
0.5 5
8.48
9.3 9
14.12
3 6 .86
1.84
10.18
7.22
1.32
15.73
12.91
8.48
4.0 8
11.33
1.25
9.42
11.71
13.87
6.0 4
16.82
3.9 5
9.7 9
8.44
6.3 4

89.00
90 .54
94 .57
86 .40
88.82
86 .88
6 2 .7 6
90 .2 7
6 6 .7 0
89 .26
76 .94
81 .08
91 .46
83 .38
96 .58
78 .14
88 .12
75 .03
84 .53
88 .99
85 .58
19.85
87 .25
92.72
81.30
54.73
87 .52
89 .15
89 .07
71.71
89.09
81.25
48 .98
96 .40
88.54
95.87
92.47
88.97
88.11
92.24
62 .07
95.85
88.75
100.00
96.85
90.63
99 .12
69.21
85.97
55.91
60 .93
98 .16
86 .54
75 .16
98 .68
53.00
86.49
65.51
95 .88
88 .2 6
98 .74
81 .23
7 5 .3 0
61 .45
90 .49
67 .04
94 .47
82.94
83.49
83.49|

0 .6 9
5.81
0 .9 4
1.47
3 .1 4
1.66
2 0 .3 2
1.31
2 4 .4 9
1.10
14 .92
5 .9 2
5 .1 0
7 .8 6
0.1 2
10 .63
3 .4 2
15 .69
0 .0 0
0 .8 0
8 .0 8
70.31
5 .2 5
0 .8 6
5 .3 0
3 0 .7 7
6 .7 6
4 .2 0
0 .4 5
2 1 .3 5
0.3 0
0.7 6
32 .9 2
1.42
0 .6 9
0.1 2
3 .1 5
3 .0 7
7.4 7
3 .9 4
25 .8 3
0.6 7
1.95
0.0 0
0.8 2
2.11
0.3 3
22.31
4 .6 4
29 .9 7
2.2 2
0.0 0
3 .2 8
17.62
0.0 0
3 1 .2 7
0.6 0
26.01
0 .0 4
0.41
0.01
9.3 5
12.98
2 4 .6 8
3.4 7
16.14
1.58
7.2 8
8 .0 8
10 .17


Figure 4.15: Percentage of each parish that has a value greater than one from the cumulative evidence.
Figure 4.15 (Cont.): Percentage of each parish that has a value greater than one from the cumulative evidence analysis for woodland.

Draycott in the Clay has by far the highest woodland cover, but this is a parish on the edge of the forest (Figure 4.14) and only part of the parish falls within the forest boundary. It is therefore not reflective of an entire parish so cannot really be compared with others.

Needwood Region has considerably more woodland than the Trent Valley, Mease Lowlands and Midlands Coalfield regions. Yet only two parishes within Needwood have potential woodland cover for more than 15% (Draycott in the Clay and Hoar Cross) so there are still considerable areas where efforts could be made to encourage planting in this region.
4.6 Discussion

4.6.1 Cumulative Evidence and LCM2000 metadata

It is important to keep in mind the datasets used in the cumulative evidence analysis, and importantly the differences between them, when interpreting the results. The datasets used here are described in detail in Section 4.2, as are the reasons for using them. Any differences are important, as the cumulative evidence analysis, as described in Section 4.3, does not give any weighting to one dataset over another. In other words each piece of evidence has equal value within the analysis, which means that it is assumed that each dataset is equally likely to accurately map the phenomenon in question. Given the need to use large extant datasets, this is not likely to be the case, however it is not felt that this weakens the results, though these factors must be considered in any interpretation.

4.6.1.1 Metadata from per-pixel classification

Figures 4.2, 4.4 and 4.6 show that, for each of the three land cover types reviewed, the metadata from the per-pixel classification gives a useful indication as to the uncertainty regarding the classification at a land parcel level. As certainty of the evidence for each land cover type increases, the percentage of the parcel designated as that category by the per-pixel classification also increases. Demonstrating that the metadata from the per-pixel classification reflects confidence from independent datasets that the category is present.

While Figures 4.2, 4.4 and 4.6 all display a similar trend in the relationship between the uncertainty analysis and the metadata from the per-pixel classification, the detail within them is quite different. In the cases of woodland and urban (Figures 4.2 and 4.4), the increase in the mean percentage is close to being linear, whereas the increase in the graph of water is not (Figure 4.6). This will be affected by the smaller sample size at higher values for water, and

122
it is possible that the relationship looks most linear in the case of urban areas because there are a greater number of samples at higher values.

For all of the three land cover types reviewed here, the per-pixel percentages of parcels intersecting areas allocated value 0 and 1 are statistically indistinguishable as the full range of the 95% confidence intervals is zero. This is still the case even at the 85% confidence level, at which point there is no overlap in the confidence intervals between any of the other values. The implication of this is that when dealing with datasets that have such heterogeneous resolutions and potential misregistrations, a single piece of evidence as to the presence of a phenomenon has to be deemed unreliable. Interestingly, in the case of woodland the smallest dataset in terms of spatial coverage, represents 2.61% of the study area (Table 4.2); at the point where more than half the datasets agree (those given a value of four or more) 2.57% of the study area is represented (Table 4.3), so the total area is very similar. This is also the case with urban areas where the smallest dataset covers 7.76% of the area and those regions of value 4 and above cover 7.43%. When only using four datasets, as with water, the area covered is not quite so similar. It is unclear if this is explained by the smaller number of datasets or the smaller areas covered, leaving potentially greater impacts from misregistration.

The graph of urban areas (Figure 4.4) is the only one to have a range for percentages of 0 to 100 for all values. The other two land cover types are more distinct within the maximum and minimum values. This is possibly due to the highly heterogeneous nature of urban areas and the per-pixel classification is designating an individual 25m pixel. However this will also be affected by the small sample size in the case of water.

Given a number of datasets with a fully shared ontology to use as inputs to the cumulative evidence analysis the results would certainly be different, as larger proportions of the study
area would have been allocated to higher values due to greater agreement between datasets. This would lead to higher sample sizes within the higher values and probably to lower average percentages for each value, though it is felt unlikely to greatly change the general pattern of the results (Figures 4.2, 4.4 and 4.6).

4.6.1.2 Metadata from per-parcel classification

It has been suggested that the spectral probability from a maximum likelihood classifier can be used as a measure of classification uncertainty (Goodchild et al., 1992; Fisher, 1994a; van Deusen, 1995; van der Wel et al., 1998; de Bruin and Gorte, 2000; Foody, 2002). Normally only the most likely class is reported, which means that not all of the information created within the classification process is recorded (Foody et al., 1992). It follows then that reporting the probability, even just that of the most likely class, as is the case with the per-parcel metadata in LCM2000, can give a measure of the spatial distribution of uncertainty (Foody, 2002). This is moving towards a soft classification approach and is how the per-parcel metadata is reported in the LCM2000, with the probability of the most likely class included.

If the per-parcel metadata from LCM2000 gives an indication of uncertainty in the classification of the parcel, then there should be a predictable relationship between it and the cumulative evidence values. Two relationships could be predicted. Firstly in those parcels that have been classified as the target land cover type the spectral probability of those parcels should increase as the certainty value increases. Secondly the opposite relationship could be predicted in parcels that were not classified as the target cover type. Figures 4.7, 4.8 and 4.9 show that no such pattern is evident. Any trends that may be present are too small to be significant. This suggests that the spectral similarity between a parcel and the training data does not indicate uncertainty in the classification (Smith and Fuller, 2002). The production
process of LCM2000 could be, at least in part, responsible for this, as the spectral probability figure relates to the average spectral response of the parcel and not of an individual pixel and so is partly a function of the segmentation algorithm used in production of the parcels. Also the reported spectral probability for the most likely land cover type relates to a land cover subclass. If these probabilities were amalgamated to a higher level of categorisation, then the relationship with the cumulative evidence values is likely to be very different.

The pattern of responses in Figures 4.7, 4.8 and 4.9 would imply that a set of input datasets with a fully shared ontology would have little impact on the relationships displayed in these graphs. Percentages display very similar ranges across all the values and therefore having larger areas allocated to higher values is unlikely to greatly alter the ranges of percentages within each value.

4.6.2 Cumulative Evidence and woodland extent in the National Forest

Taking the example of measuring woodland extent within the National Forest shows a potential application of the cumulative evidence method for land management. This work gives estimates of the level of woodland cover within the National Forest and shows the spatial variation in woodland cover within the forest. Outputs of this analysis are presented for the whole forest, for each of the landscape regions and also for each parish within the forest, showing that it can be adopted at a wide range of scales allowing comparison between regions as well as within regions by breaking them down into smaller geographical units. The example of the National Forest shows how useful this can be, particularly when there is an interest in a particular land cover type, such as wanting to increase or decrease the amount of that phenomenon.
Areas where tree planting has taken place or that are now under woodland management schemes (as defined by the woodland regions dataset provided by the National Forest Co.) were analysed with respect to the cumulative evidence analysis output. The planted areas that had been allocated values of zero or one in the cumulative evidence analysis were added to the estimates of woodland cover, to see the effect of the work carried out by the National Forest Company. Only those areas valued zero or one were used to prevent any double counting with the areas already assumed to have woodland cover. This was done for the whole forest and for the landscape regions, it was not done at the parish level as this is not currently used as a management unit by the National Forest Company, but there is no reason why it could not be done at this level if so desired.

Overall a very small proportion of the total area of the forest is given a value of six by the woodland cumulative evidence analysis and is thus virtually certain to be woodland (0.83%, Table 4.3), although a considerable area may be woodland, having been given a value of one to five (14.76%, Table 4.3). The fact that the total area that may be woodland is greater than the largest area that is identified in any one dataset (9.83%, Table 4.2) is very worthy of note, and indicates that the analysis provides new and interesting information.

4.7 Conclusion

4.7.1 Comparison of cumulative evidence with LCM2000 metadata

The LCM2000 is a pioneer of object level reporting of data quality metadata (see Section 2.3.2). While standards have specified information for database level metadata, little has been written on object-based metadata. Therefore the measures used in the LCM2000 have been developed in a rather ad hoc fashion. This analysis shows that some of the object-based
metadata (per-pixel lists) relating to attribute accuracy available with the LCM2000, at least for certain categories, gives a useful indication as to the uncertainty of classification at a land parcel level, although other measures may not (spectral probability). This means that users can take the spatial variation of uncertainty in the data into account. While it may be intuitively appropriate to exploit the spectral probability from the maximum likelihood classifier as an indicator of attribute accuracy, the work reported here suggests that this may not be the case (Smith and Fuller, 2002). Associating the results of an entirely independent classification procedure to the map at a land parcel level, gives users of the LCM2000 a tool by which they can develop an understanding of the impact of categorical uncertainty on their work.

4.7.2 Analysis of extent of woodland

The analysis outlined here has provided spatially specific estimates of woodland cover, showing the high variability across the National Forest and the impact of recent planting schemes. Such an analysis can inform the management policy within the National Forest, highlighting areas where effort should be concentrated and also assessing how well management is meeting plans laid out in the strategy for the Forest (National Forest Co., 2004). In demonstrating a strong relationship between the cumulative evidence analysis and the PPL data, at least for woodland and urban categories, it is implied that the analysis of the presence and extent of woodland could be carried out using the PPL data from LCM2000 alone. This negates the need to compile a number of datasets, which in turn circumvents the problems of datasets not having a shared ontology.
Chapter 5: Evaluating uncertainty in classification using indices of dominance

5.1 Introduction

5.1.1 Background

Object-based metadata within the LCM2000 gives a considerable amount of information about each parcel classified in the map, including the lineage of the parcel and about the classification process. Output from the pixel-based classification (the per-pixel list or PPL) is also included and relates to the separate and independent classification carried out and gives users information about potential heterogeneity within the parcel (see Section 2.3.2). Knowledge about the heterogeneity of parcels may be important in itself to some users, as it is an important aspect in some forms of land management such as for conservation. It is also important in terms of assessing uncertainty within the map product, as heterogeneity has been shown to be a contributing factor in low classification accuracy in thematic maps (Smith et al, 2002).

In this chapter the metadata is used to generate new information about each parcel that can be used to define and analyse dominance of a land cover type within parcels as well as how this affects classification uncertainty. The PPL, which gives information regarding heterogeneity within the parcels of LCM2000 will be reviewed using indices and the output of these calculations will be compared across the land cover types of the classified polygons and across regions. Spatial autocorrelation of the index values is also calculated in order to interpret the results. The relationship between index values and biodiversity is explored to assess if the metadata relating to heterogeneity of parcels can provide a surrogate way of estimating biodiversity. This is possible as the classes that LCM2000 is attempting to map
are broad habitats, which are biological habitats, so the metadata is relating to habitat heterogeneity.

The National Forest is used as a study area (see Section 1.3 for further details). This is an area of considerable size (approx. 500km²) and it has been subdivided into areas called Landscape Regions (Figure 5.1). These regions are designed to represent distinct cultural and landscape areas, though they were drawn up to form management units for the forest and are not meant to be strict biogeographic zones. While the Landscape Regions are only loosely defined regions, they are useful subdivisions of the study area allowing a more detailed review of the results across a slightly smaller spatial scale.

**Figure 5.1:** The National Forest showing the six Landscape Regions within it.
5.1.2 Indices of heterogeneity or dominance

Ecologists have been using indices to assess diversity in natural communities for many years, such as those proposed by Shannon and Weaver (1949) and Simpson (1949). Such indices can be used to describe diversity in a single statistic. The information from the PPL bears significant resemblance to species count data used by ecologists and so such indices can be used to describe it, though some caution must be employed as only the five most common land cover types in each parcel are reported and they are given as percentage figures.

In terms of assessing if the category allocated to the parcel is correct, this is more likely to be the case if a single category of land cover is found to be dominant within the parcel. Given this, it was decided that it would be more appropriate to calculate index values of dominance rather than heterogeneity. In the cases of the diversity indices mentioned above this involves slight manipulation of the equation, but does not change the emphasis of what it is calculating.

The most commonly used diversity indices are forms of a general entropy function (Legendre and Legendre, 1998). When this function is used with the power 2, as for example in Simpson’s index, it is actually a measure of concentration or dominance. In this context, this function calculates the probability that two randomly chosen pixels belong to the same land cover type. To calculate diversity using this method it would be necessary to take the concentration value from 1, this is the equation commonly referred to as Simpson’s index, but in this instance we are interested in the measure of dominance of a single category, so this step is unnecessary.

While total dominance of a single category in a parcel can and does occur, the impact of boundary pixels, falling at transitions between land cover types will make this not particularly
Chapter 5: Evaluating uncertainty in classification using indices of dominance

common. Calculating dominance statistics will not only tell the user something about the homogeneity within the parcel and so about the effects this may have on the classification, but it is also informative about the effectiveness of the algorithm that generated the parcels.

5.2 Methods

5.2.1 Indices of dominance

A number of indices to assess dominance have been reviewed and they are described below in the terms of their use in analysing land cover types within parcels of the LCM2000. A straightforward method of assessing the dominance of the most common land cover type in the parcel is to take the difference between the percentage cover of the most common land cover type and that of the second most common.

\[ I_1 = p_1 - p_2 \]  \hspace{1cm} (1)

where \( p_1 \) is the largest percentage value for the parcel and \( p_2 \) is the second largest value for the parcel. Wood and Foody (1993) suggested this difference as a measure of confidence in classification, using the output of a maximum likelihood classifier. A high difference suggests significant dominance of the most highly represented type, whereas a low figure suggests a more even distribution of types. Such evenness is likely to mean increased uncertainty in the classification.

An alternative method of assessing dominance is given in Equation 2, proposed by Simpson (1949) and, as mentioned earlier is a general entropy function:
Chapter 5: Evaluating uncertainty in classification using indices of dominance

\[ I_2 = \sum p_i^2 \]  

(2)

where \( p_i \) is the percentage of the parcel classified as the \( i \)th land cover type. One benefit of this measure is that it would give a higher weighting to parcels where the dominance of one land cover type is great, as increasing the power of the function decreases the relative weight of rare categories (Legendre and Legendre, 1998).

Another index tested was termed \( D_1 \) by O’Neill et al (1988).

\[ D_1 = \ln n + \sum_{i=1}^{m} p_i \ln p_i \]  

(3)

where \( m \) is the total number of possible land cover types and \( n \) is the number of land cover types represented in the parcel (between 1 and 5). \( \ln n \) represents the potential maximum with all present land cover types represented to an equal amount.

Equation 3 was amended in order to express the index as a value between 0 and 1 by dividing \( D_1 \) by the potential maximum (\( \ln n \)), giving:

\[ I_3 = \frac{\ln n + \sum_{i=1}^{m} p_i \ln p_i}{\ln n} \]  

(4)

Where \( n = 1 \) it was amended to \( n = 1.00001 \) in order to prevent \( \ln n = 0 \).
Chapter 5: Evaluating uncertainty in classification using indices of dominance

The final index tested was that used to calculate classification uncertainty within the soft classification functionality of the Idrisi GIS package.

\[
\text{Classification Uncertainty} = 1 - \frac{\sum_{i} p_i}{\max - \frac{1}{n}}
\]

where \(\max\) is the maximum percentage value of any land cover type in the parcel. Equation 5 gives a measure of the lack of dominance of any particular land cover type (Eastman, 1999; p81) and so gives an opposite response to the other indices used. In order to make it a similar measure of dominance as the other indices, the equation was altered to the following:

\[
I_4 = \frac{\max - \sum_{i} p_i}{\frac{1}{n}}
\]

This now gives a measure of the difference between the maximum membership value and the total dispersion of values across all land cover types as the numerator and the difference between the maximum possible membership and the even dispersion of values across all types. \(I_4\) therefore represents the maximum value in each parcel relative to the largest possible value (Eastman, 1999).

The measures outlined above are calculated for every polygon within the National Forest. Values from the pixel-based classification were normalised to total one hundred prior to the calculations being made, in order to make the heterogeneity values of all polygons directly comparable.
Land cover classes are reported at three hierarchical levels within LCM2000. The most general level categorises parcels as a broad habitat. Subclasses of those broad habitats are also reported and then variants of those subclasses at the level of greatest detail (see Section 2.3.1). For example, the arable broad habitat has three subclasses and 22 variants within it. Dominance value calculations are carried out at each classification level in order to investigate the impact of greater generalisation in the classification structure.

All parcels in the National Forest were labelled as to which Landscape Region they fell within by adding this as an attribute. The attribute table was then imported into an MS Access database along with a table detailing all adjacent parcel identities, created using the PALINFO command of ArcInfo. From these data an individual table was created for each Landscape Region and exported to a series of MS Excel spreadsheets that were used to calculate all of the dominance indices and the adjacency data used with the index values to calculate spatial autocorrelation.

### 5.2.2 Spatial autocorrelation

The level of spatial autocorrelation in the dominance values is calculated using Moran’s I:

\[
I = \frac{n \sum_i \sum_j w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2 \sum_i \sum_j w_{ij}}
\]  

(7)

where \( i \) and \( j \) refer to parcels and \( y \) is the index value of the parcel. The numerator of the second fraction is a covariance term and \( w_{ij} = 1 \) when \( i \) and \( j \) are neighbours and 0 otherwise, thus ensuring that the calculation only relates to adjacent parcels. The first fraction divides...
the second by the variance of the whole dataset in order to normalise the equation ensuring that the result is not affected by the magnitude of the values of \( y \) (Goodchild 1986; O’Sullivan and Unwin, 2003).

### 5.2.3 Comparison with biodiversity

As the calculated indices are assessing the diversity of land cover types, and therefore habitat types, within parcels of the LCM2000, they were compared with biodiversity in order to see if greater intra-parcel heterogeneity in LCM2000 meant increased biodiversity. The subject of biodiversity is a complex one and it is not intended to discuss the concept here. For the purpose of this work an estimate of the number of species recorded in an area was felt to be an appropriate measure to use. Biological records data were obtained from the National Biodiversity Network (NBN) website (www.SearchNBN.net). This data includes all records lodged with the NBN from a wide range of datasets; all those included in the analysis are listed in Table 5.1 and only include national datasets as some counties have more local datasets than others. Including local datasets would give greater emphasis to some geographical regions within the study area than others.

Only those data recorded from 1976 onwards were retained, in order to reduce the effect of historical records. It is also recognised that from the mid 1970’s onwards there has been a more consistent effort made in biological recording (S. Wilkinson, pers. comm.). The biological sample data is recorded with four levels of precision in their location information and these relate to squares in the Ordnance Survey Grid at 100m, 1,000m, 2,000m and 10,000m resolution. Only the two most precise were used as the lower resolution data was felt to be too coarse for the study area.
Chapter 5: Evaluating uncertainty in classification using indices of dominance

In the case of the 100m precision data, locations for the centre points of the 100m squares were calculated in order to use as the locations for the records. The number of different species recorded within each square was counted and this value was interpolated across the study area. The interpolation method used was Inverse Distance Weighted (IDW), using the power 2 and the nearest 20 points. In order to utilise the 1,000m precision data, they were combined with the 100m records and the relevant centre point of each 1,000m square was calculated for each one. These data were also converted into a species count and interpolated in the same way as the 100m records. Both interpolated layers were then summarised over the parcels of LCM2000 using the Zonal Statistics method in ArcInfo. The mean value of species count for each parcel could then be compared with the index values.

Table 5.1: Datasets from which biological samples were obtained via the NBN Gateway (www.SearchNBN.net).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of records dated 1976 or later</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Atlas of British and Irish Flora 2002</td>
<td>14,993</td>
</tr>
<tr>
<td>Biological Record Centre - Cranefly</td>
<td>996</td>
</tr>
<tr>
<td>Biological Record Centre – Dragonfly and Damselfly</td>
<td>1,318</td>
</tr>
<tr>
<td>Biological Record Centre – Ladybird survey</td>
<td>52</td>
</tr>
<tr>
<td>Biological Record Centre – Mammals database</td>
<td>149</td>
</tr>
<tr>
<td>Biological Record Centre – Orthoptera</td>
<td>301</td>
</tr>
<tr>
<td>Biological Record Centre – Reptiles and amphibians</td>
<td>271</td>
</tr>
<tr>
<td>Biological Record Centre – Trichoptera</td>
<td>118</td>
</tr>
<tr>
<td>British Bryological Society</td>
<td>2,627</td>
</tr>
<tr>
<td>Conchological Society</td>
<td>599</td>
</tr>
<tr>
<td>Dragonfly recording scheme</td>
<td>1,768</td>
</tr>
<tr>
<td>Ground Beetle recording scheme</td>
<td>672</td>
</tr>
<tr>
<td>Hoverfly recording scheme</td>
<td>3,562</td>
</tr>
<tr>
<td>Spider recording scheme</td>
<td>10,242</td>
</tr>
</tbody>
</table>

The method already described assumes that the biological data is full and at an appropriate resolution to compare with the LCM2000. Due to the resolution of the biological records being more coarse that the parcels of LCM2000, the index values were also summarised over
Chapter 5: Evaluating uncertainty in classification using indices of dominance

a 1000m grid using Zonal Statistics and the mean value compared with the species count for that square. This was done for all squares with biological data. The same analysis was carried out at an even more coarse resolution by summarising the data to the Landscape Region level.

5.3 Results

5.3.1 Alternative indices

The index values as calculated at the broad habitat level, using equations 2, 4 and 6 were compared to the difference statistic given by equation 1 to see if they gave more detail regarding the dominance value. These comparisons are shown in Figures 5.2, 5.3 and 5.4.

![Figure 5.2: Comparison of the dominance calculation from Simpson’s Index, $I_2$ (equation 2), with the difference statistic, $I_1$ (equation 1). Red line denotes the limit of the possible values.](image)

The curved shape of the relationship described in Figure 5.2 shows how the calculation given by equation 2 adds weight to situations where the dominance of a single land cover type is great (Legendre and Legendre, 1998). Therefore the value of $I_2$ increases at an increasing rate.
Chapter 5: Evaluating uncertainty in classification using indices of dominance

as \( I_1 \) approaches 1, which is the point of dominance by a single category. The range of values of \( I_2 \) when \( I_1 \) is low reflects the number of classes included in the calculation. The minimum value of \( I_2 \) is when all values are equal and therefore evenly distributed. Such a scenario would render a difference statistic value of \( I_1 = 0 \). However, when there are five land cover types taken into consideration the minimum value of \( I_2 \) is:

\[
I_2 = 0.2^2 + 0.2^2 + 0.2^2 + 0.2^2 + 0.2^2 = 0.2
\]  

(8)

when there are only two land cover types the minimum value is:

\[
I_2 = 0.5^2 + 0.5^2 = 0.5
\]

(9)

So the situations described in equations 8 and 9 demonstrate that while \( I_1 \) has a value of zero, \( I_2 \) can have a range of values between 0.2 and 0.5. This is an interesting feature of the calculation, but does not necessarily give a stronger impression of the dominance of a category. \( I_2 \) will also give a stronger result when there are two categories with similar values within a parcel, whereas \( I_1 \) would have a low value, which again is more appropriate to the question of dominance.

![Comparison of the dominance calculation, \( I_1 \) (adjusted from O’Neill et al., 1988; equation 4) with the difference statistic, \( I_1 \) (equation 1). Red line denotes the limit of the possible values.](image)

**Figure 5.3:** Comparison of the dominance calculation, \( I_1 \) (adjusted from O’Neill et al., 1988; equation 4) with the difference statistic, \( I_1 \) (equation 1). Red line denotes the limit of the possible values.
The relationship between \( I_1 \) and \( I_3 \) (Figure 5.3) shows a similar pattern for similar reasons and so again does not appear to improve on the difference statistic, \( I_1 \) for my purpose.

\[ I_4 \text{ (equation 6)} \]

Figure 5.4: Comparison of the dominance calculation, \( I_4 \) (adjusted from Eastman, 1999; equation 6) with the difference statistic, \( I_1 \) (equation 1). Red line denotes the limit of the possible values.

\( I_4 \) again shows a similar relationship with \( I_1 \) (Figure 5.4), though it is a more linear pattern than the previous two examples. Again it appears to pick out more detail when the value of \( I_1 \) is low but does not improve on the dominance value given by the difference statistic, \( I_1 \).

Given these findings the calculated value of \( I_1 \) is used for subsequent aspects of this study.
5.3.2 Variation of dominance across land cover types and levels of classification

Dominance values analysed as a mean of values at the broad habitat level of classification (Figure 5.5) show that the arable category clearly has the highest values demonstrating that there is little confusion within parcels between this broad habitat and others. The lowest values are found in the rough grass, calcareous grass and bracken categories, so parcels given these types appear to be more mixed according to the PPL, raising questions about the reliability of their classification.

Figure 5.5: Difference between two highest percentages ($I_1$ from equation 1) by land cover class at the broad habitat level. The numeric codes of the broad habitats are shown in brackets on the x-axis labels.
Figure 5.6: Comparison of the difference between two highest percentages ($I_1$ from equation 1) across different levels of classification grouped by broad habitat. Graph a shows arable land cover types. See Appendix 1 for all land cover codes.
Chapter 5: Evaluating uncertainty in classification using indices of dominance

Figure 5.6 (cont.): Comparison of the difference between two highest percentages ($l_i$ from equation 1) across different levels of classification grouped by broad habitat. See Appendix 1 for all land cover codes.

Mean index values are also shown across the other classification levels of subclass and variant to allow comparison (Figure 5.6). Some of the classes display considerable differences.
Chapter 5: Evaluating uncertainty in classification using indices of dominance

between the different levels. In the case of arable (Figure 5.6a) there are very variable index values across the variants, but when analysed at the broad habitat level the value is higher than any of the variants, showing that much of the confusion between variants would appear to be within the arable broad habitat.

Other major differences are found in the improved grass and urban broad habitats. The subclass 5.2 (setaside grass) is considerably lower than subclass 5.1 within improved grass (Figure 5.6b) implying that this is a more problematic category. As with the arable category though, when improved grass is analysed as a broad habitat it has a high value, so it is likely that setaside is being confused with other variants within the broad habitat.

A different pattern is displayed by calcareous grass (Figure 5.6b) and inland bare ground (Figure 5.6c), which have lower values at the broad habitat level than for some of the variants, meaning that there is confusion with other broad habitats within these parcels. This is also the case with broad habitats with generally low values such as rough grass (Figure 5.6b), deciduous and coniferous woodland and bracken (Figure 5.6d).

5.3.3 Spatial Variation

The index values calculated across the study area can also be evaluated to assess whether or not there is any spatial variation in the output. The values for the whole study area are shown at the broad habitat level (Figure 5.7) and the subclass variant level (Figure 5.8). Both figures use an equal interval colour scheme to allow comparison and it is clear that the variant level figures are slightly lower. Both maps effectively show generally higher values in the west (Needwood, see Figure 5.1) than the east (Charnwood), with other obvious clusters of high values in between, particularly within the Mease Lowlands and Midlands Coalfield regions.
Chapter 5: Evaluating uncertainty in classification using indices of dominance

Figure 5.7: Difference statistic at broad habitat level across the National Forest. Landscape region boundaries are shown with solid black line.

Figure 5.8: Difference statistic at subclass variant level across the National Forest. Landscape region boundaries are shown with solid black line.
Chapter 5: Evaluating uncertainty in classification using indices of dominance

In order to review these differences in more detail Figures 5.9, 5.10 and 5.11 show the broad habitat level index values for some of the regions at a smaller scale alongside maps of the LCM2000 showing the actual land cover types allocated to these areas (see appendix 2 for LCM2000 colour scheme). In conjunction with these maps Table 5.2 gives overall mean values of $I_1$ across the regions and Table 5.3 presents the levels of spatial autocorrelation calculated using Moran’s I (see Goodchild, 1986 or O’Sullivan and Unwin, 2003).

The Mease lowlands region (Figure 5.9a) displays a very particular pattern with high index values around the outer parts and lower values in the centre giving the region the strongest spatial autocorrelation of all the regions (Table 5.3) and relatively high index values (Table 5.2). In this case the high values are following a very similar distribution to the arable land cover type (brown shades in Figure 5.9b). Charnwood by contrast (Figure 5.10a) shows a very random distribution of values, which is supported by very weak autocorrelation (Table 5.3). It also has the lowest mean index values (Table 5.2). The scattered high values shown on the map do tend to coincide with arable and deciduous woodland (brown and red respectively in Figure 5.10b), but the pattern of land cover in this region is much more varied than in the Mease Lowlands, which has larger contiguous areas of a single land cover.

Needwood and Trent Valley also represent opposites (Figure 5.11) in that Needwood has high index values but weak autocorrelation and Trent Valley has low mean index values but relatively strong autocorrelation. They are also very different in terms of land cover with Needwood being a scattered mixture of improved grass and arable categories (bright green and brown, Figure 5.11b) and the north of Trent Valley is dominated by the urban area of Burton-upon-Trent. The higher index values that do occur in Trent Valley generally coincide with the arable areas in the south and the urban district with the lower values being in the more semi-natural categories in the region.
Chapter 5: Evaluating uncertainty in classification using indices of dominance

Table 5.2: Mean difference values ($I_1$) for the Landscape Regions within the National Forest at different levels of classification. Grey shading denotes rank order (dark = high value).

<table>
<thead>
<tr>
<th>Landscape Region</th>
<th>Broad habitat level mean $I_1$</th>
<th>Subclass level mean $I_1$</th>
<th>Variant level mean $I_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calke Uplands</td>
<td>52.926</td>
<td>47.876</td>
<td>45.666</td>
</tr>
<tr>
<td>Charnwood</td>
<td>44.864</td>
<td>42.197</td>
<td>40.634</td>
</tr>
<tr>
<td>Mease Lowlands</td>
<td>60.873</td>
<td>54.318</td>
<td>48.767</td>
</tr>
<tr>
<td>Midlands Coalfield</td>
<td>53.538</td>
<td>47.092</td>
<td>44.604</td>
</tr>
<tr>
<td>Needwood</td>
<td>59.107</td>
<td>55.449</td>
<td>53.065</td>
</tr>
<tr>
<td>Trent Valley</td>
<td>51.639</td>
<td>44.668</td>
<td>39.926</td>
</tr>
</tbody>
</table>

Table 5.3: Spatial autocorrelation in difference values ($I_1$) measured using Moran’s I for the Landscape Regions within the National Forest.

<table>
<thead>
<tr>
<th>Landscape Region</th>
<th>Broad habitat level Moran’s I</th>
<th>Subclass level Moran’s I</th>
<th>Variant level Moran’s I</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calke Uplands</td>
<td>0.096</td>
<td>0.129</td>
<td>0.162</td>
</tr>
<tr>
<td>Charnwood</td>
<td>0.066</td>
<td>0.082</td>
<td>0.098</td>
</tr>
<tr>
<td>Mease Lowlands</td>
<td>0.145</td>
<td>0.203</td>
<td>0.335</td>
</tr>
<tr>
<td>Midlands Coalfield</td>
<td>0.113</td>
<td>0.175</td>
<td>0.224</td>
</tr>
<tr>
<td>Needwood</td>
<td>0.071</td>
<td>0.094</td>
<td>0.119</td>
</tr>
<tr>
<td>Trent Valley</td>
<td>0.120</td>
<td>0.197</td>
<td>0.326</td>
</tr>
</tbody>
</table>
Figure 5.9: Broad Habitat level difference statistic (a) and original LCM2000 (b) for the area of the Mease Lowlands Landscape Region (black boundary line). Legend for map a is the same as in Figure 5.6, see Appendix 2 for LCM2000 colour scheme shown in map (b)
Figure 5.10: Broad Habitat level difference statistic (a) and original LCM2000 (b) for the area of the Charnwood Landscape Region (black boundary line). Legend for map (a) is the same as in Figure 5.6, see Appendix 2 for LCM2000 colour scheme shown in map (b).
Figure 5.11: Broad Habitat level difference statistic (a) and original LCM2000 (b) for the area of the Needwood and Trent Valley Landscape Regions (black boundary lines). Legend for map (a) is the same as in Figure 5.6, see Appendix 2 for LCM2000 colour scheme shown in map (b).
There are also differences in the mean index values across the Landscape Regions for individual broad habitats (Table 5.4). Of particular note, several of the land cover types have a much higher mean index value in a single Landscape Region. Examples of this are deciduous woodland in Charnwood, coniferous woodland in the Calke Uplands, arable in the Mease Lowlands, improved grass in Needwood and urban in Trent Valley. In all of these examples it is the case that this is the region where that land cover type occurs most densely and in contiguous areas. A similar pattern is observable for calcareous grass, but it is less pronounced. The remaining categories, rough grass, fen, bracken, open water and bare ground all have relatively few parcels in the study area and so it is not possible to discern if the pattern is followed.

Table 5.4: Relative proportions of $I_1$ values across the Landscape Regions compared with the average value for the entire National Forest. Those more than 10% above the average are shaded red and those more than 10% below the average are shaded blue.

<table>
<thead>
<tr>
<th>Cover type</th>
<th>Charnwood</th>
<th>Calke Uplands</th>
<th>Midlands Coalfield</th>
<th>Mease Lowlands</th>
<th>Needwood</th>
<th>Trent Valley</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deciduous woodland (1)</td>
<td>1.127</td>
<td>1.086</td>
<td>0.903</td>
<td>1.017</td>
<td>0.848</td>
<td>0.811</td>
</tr>
<tr>
<td>Coniferous woodland (2)</td>
<td>1.085</td>
<td>1.334</td>
<td>0.737</td>
<td>0.651</td>
<td>1.075</td>
<td>0.656</td>
</tr>
<tr>
<td>Arable (4)</td>
<td>0.845</td>
<td>1.000</td>
<td>1.040</td>
<td>1.124</td>
<td>0.888</td>
<td>0.934</td>
</tr>
<tr>
<td>Grass improved (5)</td>
<td>0.808</td>
<td>0.853</td>
<td>0.896</td>
<td>0.977</td>
<td>1.296</td>
<td>1.010</td>
</tr>
<tr>
<td>Grass rough (6)</td>
<td>0.875</td>
<td>0.871</td>
<td>1.062</td>
<td>0.792</td>
<td>0.000</td>
<td>1.081</td>
</tr>
<tr>
<td>Grass calcareous (7)</td>
<td>1.027</td>
<td>1.054</td>
<td>0.980</td>
<td>1.013</td>
<td>0.990</td>
<td>0.888</td>
</tr>
<tr>
<td>Bracken (9)</td>
<td>0.720</td>
<td>1.522</td>
<td>1.243</td>
<td>0.000</td>
<td>0.000</td>
<td>1.674</td>
</tr>
<tr>
<td>Fen, marsh, swamp (11)</td>
<td>0.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Open water (13)</td>
<td>1.167</td>
<td>1.103</td>
<td>0.805</td>
<td>0.605</td>
<td>0.570</td>
<td>1.054</td>
</tr>
<tr>
<td>Inland bare ground (16)</td>
<td>1.273</td>
<td>1.110</td>
<td>0.878</td>
<td>0.923</td>
<td>0.146</td>
<td>1.046</td>
</tr>
<tr>
<td>Urban/suburban (17)</td>
<td>0.897</td>
<td>0.842</td>
<td>1.040</td>
<td>0.851</td>
<td>0.722</td>
<td>1.115</td>
</tr>
</tbody>
</table>

150
5.3.4 Potential relationship with biodiversity

To reliably compare the LCM2000 with biodiversity data, the latter would need to be continuous and at high resolution. As can be seen from Figure 5.12, the data available from national datasets is far from complete even when considered at the 1km square level of precision. From the distribution of points shown it would appear that the level of survey effort has been greater in the east of the National Forest than in the west, which would mean it more likely that high species counts are achieved in the east, which is indeed the case.

The general pattern of high species counts in the east and lower in the west is exactly opposite to the general pattern of dominance index values, which display high dominance of single land cover categories in the west (Figures 5.7, 5.8 and 5.13). Despite this similarity in general pattern, there is very weak correlation with the species count values when summarised across the parcels of the LCM2000. This result may relate to insufficient or patchy biological data, or to the resolution of the biological data being inappropriate for comparison with the LCM2000. As a consequence the species count data was summarised at the broader, landscape region level (Table 5.5). The relationship between the species count and the broad habitat level index values summarised at this scale displayed only a moderate negative correlation (-0.524 Pearson Correlation Co-efficient, \( n = 6, \ p > 0.05 \)). Charnwood has the highest mean species count of all the regions (Table 5.5) and the lowest dominance index values (Table 5.2), which corresponds with the expected relationship. Calke Uplands, Midlands Coalfield and Needwood also follow the same pattern, but the Mease Lowlands and Trent Valley regions do not. Mease Lowlands has high dominance and a relatively high mean species count; though this result is heavily biased by two squares with very high species counts in the east of the region (Figure 5.12). The rest of the region has much lower values and the two higher squares may well relate to higher sampling effort. Trent Valley on the other hand has very low species counts and low overall dominance values (Table 5.5), but it is
an area of very mixed dominance values. The urban area in the north has relatively high
dominance values in relation to other urban areas (Table 5.4) and low species count (Figure
5.12), so this part of the region fits the expected pattern. South of the town though, where
dominance values are low there are also low species counts. This may again relate to the
level of sampling effort, or possibly to incomplete data across the region failing to give a
complete picture of the biodiversity.

<table>
<thead>
<tr>
<th>Landscape Region</th>
<th>Mean species count</th>
<th>Mean index value ($i_d$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calke Uplands</td>
<td>7.123</td>
<td>52.93</td>
</tr>
<tr>
<td>Charnwood</td>
<td>10.518</td>
<td>44.86</td>
</tr>
<tr>
<td>Mease Lowlands</td>
<td>7.360</td>
<td>60.87</td>
</tr>
<tr>
<td>Midlands Coalfield</td>
<td>6.177</td>
<td>53.54</td>
</tr>
<tr>
<td>Needwood</td>
<td>4.774</td>
<td>59.11</td>
</tr>
<tr>
<td>Trent Valley</td>
<td>4.071</td>
<td>51.64</td>
</tr>
</tbody>
</table>

To analyse the relationship at a scale that is relevant to the biological sample data, the index
values were summarised to the 1km squares that the species count data is reported in (Figure
5.13). The mean index values were compared with the species data but again showed a very
weak correlation ($r = -0.110, n = 306, p > 0.05$).
Chapter 5: Evaluating uncertainty in classification using indices of dominance

**Figure 5.12:** Count of recorded species for 1km squares where sample data is available

**Figure 5.13:** Mean index values ($I_1$) for 1km squares across the National Forest calculated at the broad habitat level
5.4  Discussion

5.4.1  Alternative indices

All four indices investigated here were capable of reflecting the level of dominance of the most widespread land cover type within each parcel. There are subtle differences between them in the level of emphasis given to more rare land cover types. $I_3$ is probably not suitable in this case because of the fixed number of possible categories due to only the five most common categories being reported in the metadata and also because of the need to manipulate the value of $n$ in the case of $n = 1$. It was decided to use just one index as the point of the work was to assess the effects of dominance on uncertainty in the LCM2000 and assess its relationship with biodiversity, rather than the impact of the choice of index on the analysis. Given that three indices were deemed to be suitable, $I_1$ was chosen, as it is the least complex to understand and so easiest to interpret the results.

5.4.2  Variation of dominance across land cover types and levels of classification

There is considerable variation in the level of dominance displayed in parcels of different land cover types. This is an expected response as some land cover types are easier to distinguish with the techniques used to create LCM2000 than others. This also relates to the ability of the segmentation algorithm to detect appropriate parcels. Some will be more heterogeneous because of the nature of the land cover in the area and the restrictions placed on the algorithm relating to the required size of parcels. Some areas will have smaller patches of land cover, but they may be more difficult to pick out based on the spectral response recorded in the satellite imagery. These inevitable problems will be reflected in the amount of mixed parcels and some land cover types will be more likely to suffer from these problems than others.
Chapter 5: Evaluating uncertainty in classification using indices of dominance

The index analysis also picks out the fact that some variants have quite low dominance values, but by analysing across the different levels of classification it can be seen that much of the confusion is between variants within a broad habitat. Foody (2002) notes that misclassification between similar classes may be of less significance than misclassifications between more disparate classes.

5.4.3 Spatial variation

Very clear spatial patterns are apparent in the index values (Figures 5.9, 5.10 and 5.11). Some areas have low or high values because of the land cover types that occur in those areas and some land cover types have lower mean values than others, as already discussed (Figure 5.5). It is also clear however that there are more complex interactions taking place. The existence of large contiguous areas of a land cover type appears to increase the dominance index values for the area. Mease Lowlands is an example of this; it is a region that has areas dominated by arable land cover (over 55%) forming large blocks of land. Arable has the highest index values of all the land cover types so it is unsurprising that the high values in Mease Lowlands follow the distribution of the arable areas around the edges (Figure 5.9). But the arable index values in this region are higher than the average value for arable types across the whole forest (Table 5.4).

Charnwood has the lowest index values and has a very high proportion of calcareous grass (around 20%), which has a low average index value across the whole study area (Figure 5.5). It also has a high proportion of arable (around 25%), but the arable polygons in this region have a much lower index value on average than for the whole study region, quite the opposite to arable parcels in the Mease Lowlands (Table 5.4). The clear difference between them is
that the arable in Charnwood is much more scattered and not in contiguous or clustered areas, as they are in Mease Lowlands.

The Needwood and Trent Valley regions are more difficult to interpret. Needwood has generally high dominance values explained by the fact that it is dominated by improved grass, which as a land cover type has high values (Figure 5.5), and it has higher dominance values in this region than in the study area as a whole (Table 5.4). However very weak spatial autocorrelation is displayed (Moran’s I very close to zero, Table 5.3). This could be explained by the scattered pattern of arable areas throughout the grassland, which have an average index value much lower than the study area mean (Table 5.4).

Trent Valley has slightly lower than average dominance values on the whole (Table 5.2), but does display strong autocorrelation in the values (2nd highest of the regions, Table 5.3), which appears to be explained by the large urban areas (Burton-upon-Trent). Urban parcels have higher dominance values on average in this region than for the study area (Table 5.4) and are clustered mainly in the north part of the region. This also fits the pattern of contiguous areas having high dominance values.

5.4.4 Potential relationship with biodiversity

Though several methods were attempted to assess the relationship between the index values and species count data, no demonstrable relationship was found. Different methods were pursued in an attempt to ensure that the most relevant scale was used in the analysis. While no statistically significant relationship could be shown, the maps of the two datasets display a strong similarity at a broad scale. If it is assumed that a broad scale relationship does exist then some of the observations drawn in Section 5.4.3 can be further explored in this context.
The relatively high dominance values of arable parcels in the Mease Lowlands do coincide with low species counts and lower relative dominance values for arable parcels in Charnwood also coincide with high species count data (Table 5.4 and Figure 5.12). Survey effort is a problem here as the data is incomplete in the Mease Lowlands area, but the descriptions of regions from the landscape character work carried out by the National Forest Company do relate to these observations (NFC, 2004). Areas of intensive cropping, such as the case in the Lowlands are noted as having large, open fields with only occasional and poorly defined hedgerows. Charnwood on the other hand is noted as having hedged fields in lower areas and hedgerows are widely considered to be very important features for increasing biodiversity. The scattered nature of the arable fields in Charnwood also mean that they are interspersed with other land cover types and therefore habitats. This means that they are likely to be associated with diverse habitats such as ancient woodland (NFC, 2004), which tend to be very diverse habitats. Being within a more diverse landscape, as Charnwood is in comparison with Mease Lowlands, is likely to lead to there being a larger species pool (A. Gray, pers. comm.), providing more biodiversity across the region.

Urban areas are less well understood in terms of their biodiversity, largely as they have been studied less than rural areas. Their importance for biodiversity however is becoming more apparent as understanding increases, so urban areas are not necessarily wastelands in terms of biodiversity, but more diverse urban landscapes lead to greater opportunity for wildlife. As with the arable example, the contiguous urban areas of the large town of Burton-upon-Trent had relatively higher dominance values than the smaller settlements across the National Forest (Table 5.4). As a larger town it has more industrial areas and more dense housing, which means smaller gardens. These factors will lead to lower biodiversity, so again this fits the broad pattern of relationship between dominance and biodiversity.
5.5 Conclusion

Object-level metadata within the LCM2000 is very detailed and may prove daunting to some users. Using an index to reduce the PPL element of it to a single statistic makes it easier to use, though it is important that the index is easy to understand and giving a clear description in order for it to be interpreted correctly.

Utilising an index of dominance allows the user to draw conclusions regarding the likely effects of mixed parcels on classification uncertainty. It has also been shown here to give new and useful information about which broad habitats are most affected by such heterogeneity and whether any confusion within parcels is between variants of one broad habitat or between different broad habitats, which is likely to be a greater problem.

Spatial variation in the dominance is shown to follow distinct patterns following the pattern of land cover itself, as some types have higher dominance than others. But it is also strongly affected by the distribution of the land cover types, as a more scattered distribution of a land cover will lead to lower dominance values than where it occurs in large contiguous areas.

Broad patterns of dominance do appear to display an inverse relationship with species count data, but this is not a strong relationship and no firm conclusions can be drawn from this analysis. Part of the difficulty in assessing this relationship is that the biological records are variable across space and the level of sampling effort is not consistent. These problems would need to be taken into account in order to review this further. It was not possible to do so here due to time constraints, but due to the apparent broad scale relationship that has been noted, this would be an important area of further work to locate more biological data and address the problems of scale and survey effort.
Chapter 6: Attribute measurement within fixed land parcels and grown regions

6.1 Introduction

The work presented in this chapter measures the level of support for land cover types within the metadata of the Per Pixel List (PPL). A theoretical framework for this measurement has already been outlined in terms of the types of data being used (Section 2.4); here that framework is extended to describe the way these data types relate to geographical data. Initially the attribute measurement is carried out within each fixed land parcel and the support for the classification of the parcel is assessed. Controlling the spatial component of the data then allows the expansion of the measurement beyond individual land parcels. By following these two processes it is possible to detect locations where uncertainty is high and also where alternative classifications may be more appropriate.

6.2 LCM2000 metadata within theoretical framework

As discussed in Section 2.4, land cover maps are categorical data and as such this work can be viewed within the framework of scales of measurement proposed by Stevens (1946). The values in the PPL provide a membership function for a nominal category and the work reported below uses the PPL data in this way, analysing values to establish whether or not the data in the PPL meet a criteria in order for the parcel to be assigned to a category or not. Molenaar (1998; p138) specifies the function by which this can be done. The criteria used here to make this decision are in the form of a percentage cover threshold.

In analysing this data it would have been possible to create areas of similar response from the per-pixel classification if the whole raster output had been reported, but it was only summarised per parcel in the form of the PPL. This means that the attribute measurement can
only be carried out within multiples of the fixed areas of the predefined land parcels. The first 
stage of analysis reported here takes that measurement as it is reported for individual parcels 
in the PPL. Measurement of the attributes within merged adjacent parcels is then analysed, 
thus using the space component as the control (see Section 2.4), with all measured values 
from the PPL being assessed against predefined descriptions of different land covers.

The metadata within the PPL gives detail regarding the heterogeneity of each land parcel (see 
Chapter 5) and also the different land cover types that are likely to occur within it, as detected 
by the per pixel classification. Further to this, it relates to a smaller scale than the parcel level 
classification and so is also believed to give information relating to small groups of trees or 
individual buildings as the pixels are just 25m. Given this the PPL can be considered as a 
surrogate for percentage cover of land cover types within the parcel. Although in order to do 
this an assumption must be made that if a pixel is designated as the target land cover type then 
it is entirely covered by, or is 100%, that land cover type (Robinson et al, 2004). This is not 
an unreasonable assumption when dealing with 25m pixels, though the percentage could 
easily be varied in the calculations (see Equation 6.1) in order to soften this assumption. This 
type of percentage cover information coincides with the form taken by some definitions of 
land cover types. An example of this is the definition of what constitutes woodland within the 
Countryside Survey 2000, being areas with at least 20% cover by tree canopies, where the 
trees are greater than 5m tall (Briggs, 2003).

The questions therefore arise whether all the parcels classified as a particular land cover type 
meet the definition criteria within their PPL for that type and if not, what cover types are 
supported? Are the same cover types consistently supported in these cases and does the lack 
of support from the PPL point to misclassification? Also, are there other parcels that meet the 
criteria, even though they are not classified as that cover type?
The National Forest is again used as the study area for this work and the categories selected to assess these questions are woodland and urban. The former was selected as increasing the level of woodland cover is the primary objective of the National Forest Company, the latter because urbanisation is a significant issue with regards to planning and land management. In the case of woodland many definitions are based on land cover of tree canopies and so makes an ideal example for this analysis. As far as the author is aware, all the land cover based definitions of what constitutes an urban area rely on the density of buildings (see ODPM, 2002 for a description). Such land cover information is not available within the LCM2000 and would be difficult to derive accurately from 25m resolution satellite imagery. So a land cover based definition of urban areas that could be used in the same way as that for woodland had to be created.

The vector parcels within LCM2000 were created from the original satellite imagery using a segmentation algorithm (Fuller et al, 2002b). As such they should theoretically relate to specific geographical phenomena with distinct spectral characteristics, but in practice this is not necessarily the case. There are a number of reasons for this, including problems with the original image, locations of seed points created by the algorithm, complexity of the landscape and land cover change. Any one, or a combination of such issues, can mean that parcels do not relate to specific objects. It is possible therefore that even if a parcel did not meet the criteria based on its individual metadata, it could meet the relevant criteria if it were considered as part of a larger area, having been combined with a neighbour.

It is also likely that the metadata could be used in a similar approach, to inform algorithms for mapping sub-pixel boundaries, such as those devised by Atkinson (1997). This could be
useful if there was a need to break down the parcels into smaller areas, but is not explored here, as it is not part of the aim of this work.

6.3 Methods

6.3.1 Definitions

6.3.1.1 Woodland and multiple land cover types

The work reported here uses land cover-based definitions of target land cover types and the metadata within LCM2000 to assess the uncertainty of classification. In the case of the woodland land cover type the National Forest Company, the organisation established to implement the strategy for the area, use the definition of woodland as an area with at least 50% tree canopy cover (The National Forest Co., 2004). A different definition of woodland is utilised by the Countryside Survey 2000, of which the LCM2000 is a part, which requires at least 20% tree canopy cover (Briggs, 2003).

These examples show how land cover categories can be defined in many different ways. The per pixel classification can give an indication of all land parcels that meet such criteria, even if those parcels are not classified as the target cover type within the database. Such cases would represent a potential increase in that cover type from the original classification and the converse could also be true, where a parcel is classified as the target cover type does not meet the specified criteria.

The approach of assessing the make up of the parcel, as described in the metadata, against a definition of a land cover type, can be adopted for any land cover type that can be clearly defined in a similar way to those outlined here. By doing this exhaustively, some areas would fulfil the criteria for more than one land cover type, leading to a need to define multiple land
cover categories, urban-woodland for example. As some urban areas could easily meet the criteria of 20% tree cover, in fact planning guidelines laid out in the Strategy of the National Forest Company specify that all new residential developments over 0.5ha and other developments greater than 1ha must include 20% of the area to be woodland planting (The National Forest Co., 2004). Such multiple categories do not fit into a traditional Boolean model, which implies that all locations can be mapped as exhaustive and mutually exclusive sets, as this model does not account for the vagueness involved in geographical objects (Zhang and Goodchild, 2002, p91). They are however straightforward to deal with using a multiset approach (Fisher, 2001). In the example of urban-woodland some areas, whilst clearly urban, would also meet the criteria of 20% tree canopy cover. Such treatment can therefore assist in describing the heterogeneous nature of land cover exhibited in some areas and could be argued to give a more accurate description of the landscape.

Woodland cover types are defined in many ways relating to the percentage cover of tree canopies as has already been mentioned. Threshold values for this land cover type are therefore straightforward to select with the Countryside Survey 2000 using a threshold of 20% and the National Forest adopting 50% (Table 6.2).

6.3.1.2 Urban land cover types

The urban category is more problematic as the urban class does not have land cover-based definitions in use such as those used for woodland. Definitions of what constitutes an urban area can relate to a number of different approaches. Examples put forward in the User Guide to Urban and Rural Area Definitions published by the Office of the Deputy Prime Minister (ODPM, 2002) include three types of approach based on land use, functional area or density of population or buildings. The definition used in the Urban Settlement boundary dataset published by the ODPM (see Chapter 4 for further details of this dataset) claims to utilise a
land use approach, though it is actually based solely on population density within artificial boundaries used in the UK census. Urban areas were extracted from 1:10,000 OS data, each area had to include four 1991 census enumeration districts (ED’s) and have a population of at least 1,000 (ODPM, 2002). Obviously such definitions do not relate purely to land cover and require subsidiary information, in this case from the census, so are not comparable to the woodland examples above.

Table 6.1: Analysis of OS Mastermap attribute data for all areas within the National Forest defined as Urban by the ODPM Urban Settlements dataset

<table>
<thead>
<tr>
<th>Surface description</th>
<th>Percentage of urban areas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manmade surface (inc. building, road and railway)</td>
<td>33.61</td>
</tr>
<tr>
<td>Multiple surface (garden)</td>
<td>40.08</td>
</tr>
<tr>
<td>Natural surface (inc. trees grass and scrub)</td>
<td>25.09</td>
</tr>
<tr>
<td>Rock/Unknown</td>
<td>0.98</td>
</tr>
<tr>
<td>Water</td>
<td>0.23</td>
</tr>
</tbody>
</table>

As no definition of urban areas based purely on land cover composition was available, it was necessary to describe such areas in order to derive that definition. Given that the per pixel classification uses only 25m pixels and that urban areas can be extremely heterogeneous by their nature, the PPL of urban, and particularly suburban, parcels is likely to contain a mixture of land cover types. It is necessary therefore to derive some indication of what comprises an urban area. In order to analyse the likely composition of an urban area, OS Mastermap topographic layer was clipped using the Urban Settlements dataset from ODPM. The areas within urban settlements were analysed using the “Make” attribute data provided within Mastermap (see Table 6.1). From this analysis it is clear that just under 34% of the land cover within these areas is unequivocally manmade and 40% is a complex mix of land cover types, given the description “multiple surface”, which includes such areas as gardens (OS, 2004: 164).
Chapter 6: Attribute measurement within parcels and grown regions

p65). Thresholds were therefore used in the subsequent analysis at slightly rounded figures requiring 35% and 75% urban land cover within the PPL of each parcel, with a third threshold of 55%, exactly halfway between the two, which allows for multiple surfaces to be classified in an even split between urban and non-urban (Table 6.2).

The “Make” attribute is effectively very similar to elements of the vegetation-impervious surface-soil (V-I-S) model first proposed by Ridd (1995) as a tool for analysing land cover and land use in urban environments. By assessing land cover in categories of manmade surfaces, natural surfaces and a mixture of the two, it follows a very similar principle to the V-I axis in the V-I-S model. The thresholds of 35, 55 and 75% therefore bear comparison to the descriptions of low-density, medium-density and high-density residential categories as described in Ridd’s model (Figure 6.1).

![Figure 6.1: Vegetation-Impervious surface-Soil (V-I-S) model adapted from Ridd (1995), Hung (2002) and Wu and Murray (2003)](image)
When reviewing the support given by the PPL to the classification within LCM2000 it is also important to assess those areas where the PPL conflicts with the classification of the parcel. What can we tell from these situations? In the case that a parcel is classified as a land cover type and the PPL does not lend support to that classification, what land cover types are supported by the PPL?

It was decided to consider the potential of these techniques by looking purely at the possibility of the LCM2000 having underestimated the land cover types being reviewed. The alternative situation, that of these cover types being overestimated, is also possible and could be similarly analysed using these techniques.

<table>
<thead>
<tr>
<th>Criteria to be investigated</th>
<th>Percentage cover</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Woodland</strong></td>
<td></td>
</tr>
<tr>
<td>CS2000 definition</td>
<td>20%</td>
</tr>
<tr>
<td>National Forest Company definition</td>
<td>50%</td>
</tr>
<tr>
<td><strong>Urban</strong> (from OS Mastermap analysis)</td>
<td></td>
</tr>
<tr>
<td>Manmade surface only regarded as urban</td>
<td>35%</td>
</tr>
<tr>
<td>Manmade surface and half of multiple surface regarded as urban</td>
<td>55%</td>
</tr>
<tr>
<td>Manmade surface and multiple surface regarded as urban</td>
<td>75%</td>
</tr>
</tbody>
</table>

6.3.2 Parcels with conflicting Per Pixel List and parcel classification

All parcels that meet the PPL criteria for the target land cover types of woodland or urban were identified and mapped (Figures 6.4 and 6.9). The parcels that meet these criteria but are not classified as the target cover type, and vice versa, are potentially of interest. Those that are classified as woodland but do not meet the woodland PPL criteria have little or no support
for the classification from the PPL (Figure 6.7), likewise for the urban category (Figure 6.10).

So what land cover type does the PPL support? Parcels that are not classified as woodland but do meet the woodland PPL criteria are areas representing a potential underestimation of woodland within the LCM2000. Parcel SK146457 is highlighted in Figure 6.2 as an example of this as the PPL contains 46% woodland.

6.3.3 Region growing methodology

A region growing method was used to assess whether the per pixel metadata for a new parcel met the criteria for woodland or urban when it was created by merging neighbouring parcels. This involves using the metadata to generate new information about a locality, thereby reusing the metadata provided by the producer. A constraint was introduced ensuring that there must be some evidence within the per pixel classification that the target cover type is present within both of the merged parcels for the new larger parcel to be considered as potentially that land cover type. This greatly reduced such problems as parcels being considered woodland just because they were adjacent to dense woodland and so inheriting high woodland per pixel values from the neighbour, even though there was no evidence in the parcel itself.

A table containing all pairs of polygons that share a boundary was created using the PALINFO command in ArcInfo. This table, along with another containing all the PPL data and areas for each polygon, was imported into a relational database, MS Access. For ease of description the methods will be described using the woodland example, although the same techniques were also used for urban land cover. Calculations were carried out to give a total percentage that was allocated to a woodland land cover type for each polygon. From this data
it is possible to calculate the proportion of each pair of polygons that would be within a woodland category (equation 6.1).

\[
\text{MPW} = \frac{(P_1 \text{ Area} \times P_1 \text{ PW}) + (P_2 \text{ Area} \times P_2 \text{ PW})}{(P_1 \text{ Area} + P_2 \text{ Area})}
\]

where MPW is the proportion of potential woodland in the merged parcel, \( P_1 \) and \( P_2 \) are the two neighbouring parcels and PW is the value from the PPL that is associated with woodland categories, expressed as a value between zero and one. Comparison of MPW with the threshold follows the function laid out by Molenaar (1998; p138).

Using the MPW statistic, merged parcels that meet the criteria of any given definition of woodland can be selected. When deciding whether or not a parcel represents a potential increase in woodland it is important to know whether or not it was classified as woodland in the LCM2000 (see Table 6.3). Sixteen possible scenarios are identified in Table 6.3, with one of three possible effects on the level of potential woodland shown in the penultimate column, being an increase, a decrease or no effect. The scenarios of interest are those in which an increase of woodland occurs, which are numbered 5, 7, 8 and 13 in Table 6.3. Scenario number 10 also creates an increase but is effectively the reciprocal of scenario number 5 and so would create double counting of parcels if it were taken into account. Of particular interest is scenario 13, in which two non-woodland parcels are merged yet still meet the criteria of being woodland. This is more likely to happen if the criteria threshold is low and should be unlikely at best when using a high threshold of 50% or greater. The parcel classification in the LCM2000 however uses the averaged spectral response of the parcel and is not related to the data in the PPL, so it is a possibility that cannot be ignored.

Figure 6.2 shows an example of how the region growing method works. Part of the detailed metadata are shown for two parcels. One is classified as woodland by the parcel classification
Chapter 6: Attribute measurement within parcels and grown regions

(SK040868) and the other was classified as calcareous grassland (SK146529). When these two parcels are merged, the calculation given in Equation 6.1 is made:

\[
\text{MPW} = \frac{(45.625 \times 0.92) + (31.875 \times 0.02)}{45.625 + 31.875}
\]

\[
\text{MPW} = 0.55
\]  

An MPW value of 55% means that this new merged parcel meets the most stringent criteria being used to define woodland and so SK146529 can be considered to be a potential increase in the amount of woodland. In Table 6.3 this example would be of type 7 as the percentage of woodland for the merged parcel (MPW) is less than that of parcel 1 in the calculation, but is still greater than 50%. This is slightly simplified for the sake of example as the PPL values for all parcels were normalised to make them all represent 100% as this is not always the case.
Figure 6.2: Example of region growing methodology. Inset shows location on OS 1:50000 mapping.
Table 6.3: Possible scenarios in region growing analysis using an example of a woodland analysis with a definition of woodland as 50% tree cover

<table>
<thead>
<tr>
<th>No.</th>
<th>Original parcel ($P_1$)</th>
<th>Adjacent parcel ($P_2$)</th>
<th>Effect of merge on PPL % for woodland</th>
<th>Effect on overall area classified as woodland</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Woodland</td>
<td>Woodland</td>
<td>↑ (&gt;50%)</td>
<td>←</td>
<td>Remains woodland</td>
</tr>
<tr>
<td>2</td>
<td>Woodland</td>
<td>Woodland</td>
<td>↓ (&gt;50%)</td>
<td>←</td>
<td>Remains woodland</td>
</tr>
<tr>
<td>3</td>
<td>Woodland</td>
<td>Non-woodland</td>
<td>↓ (&lt;50%)</td>
<td>↓</td>
<td>No longer woodland (should not happen)</td>
</tr>
<tr>
<td>4</td>
<td>Woodland</td>
<td>Non-woodland</td>
<td>← (&gt;50%)</td>
<td>←</td>
<td>Remains woodland</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No.</th>
<th>Original parcel ($P_1$)</th>
<th>Adjacent parcel ($P_2$)</th>
<th>Effect of merge on PPL % for woodland</th>
<th>Effect on overall area classified as woodland</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Woodland</td>
<td>Woodland</td>
<td>↑ (&gt;50%)</td>
<td>↑</td>
<td>Whole becomes woodland</td>
</tr>
<tr>
<td>6</td>
<td>Woodland</td>
<td>Non-woodland</td>
<td>↓ (&lt;50%)</td>
<td>↓</td>
<td>Whole no longer woodland</td>
</tr>
<tr>
<td>7</td>
<td>Woodland</td>
<td>Non-woodland</td>
<td>↓ (&gt;50%)</td>
<td>↑</td>
<td>Whole becomes woodland</td>
</tr>
<tr>
<td>8</td>
<td>Woodland</td>
<td>Non-woodland</td>
<td>← (&gt;50%)</td>
<td>↑</td>
<td>Whole becomes woodland (should not happen)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No.</th>
<th>Original parcel ($P_1$)</th>
<th>Adjacent parcel ($P_2$)</th>
<th>Effect of merge on PPL % for woodland</th>
<th>Effect on overall area classified as woodland</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>Non-woodland</td>
<td>Woodland</td>
<td>↑ (&lt;50%)</td>
<td>↓</td>
<td>Remains non-woodland</td>
</tr>
<tr>
<td>10</td>
<td>Non-woodland</td>
<td>Woodland</td>
<td>↑ (&gt;50%)</td>
<td>↑</td>
<td>Whole becomes woodland (reciprocal of No. 5)</td>
</tr>
<tr>
<td>11</td>
<td>Non-woodland</td>
<td>Non-woodland</td>
<td>↓ (&lt;50%)</td>
<td>↓</td>
<td>Remains non-woodland (should not happen)</td>
</tr>
<tr>
<td>12</td>
<td>Non-woodland</td>
<td>Non-woodland</td>
<td>← (&lt;50%)</td>
<td>↓</td>
<td>Remains non-woodland (should not happen)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No.</th>
<th>Original parcel ($P_1$)</th>
<th>Adjacent parcel ($P_2$)</th>
<th>Effect of merge on PPL % for woodland</th>
<th>Effect on overall area classified as woodland</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>Non-woodland</td>
<td>Non-woodland</td>
<td>↑ (&gt;50%)</td>
<td>↑↑</td>
<td>Whole becomes woodland (should not happen)</td>
</tr>
<tr>
<td>14</td>
<td>Non-woodland</td>
<td>Non-woodland</td>
<td>↑ (&lt;50%)</td>
<td>←</td>
<td>Remains non-woodland</td>
</tr>
<tr>
<td>15</td>
<td>Non-woodland</td>
<td>Non-woodland</td>
<td>↓ (&lt;50%)</td>
<td>←</td>
<td>Remains non-woodland</td>
</tr>
<tr>
<td>16</td>
<td>← (&lt;50%)</td>
<td>←</td>
<td>Remains non-woodland</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Having selected the pairings that represent a potential increase in woodland the $P_2$ parcels in each one are filtered to ensure that they contain some evidence of woodland in the PPL. If they contain no such evidence then they are discarded, if they do then they are considered to be potential increases in the overall amount of woodland. This step is important to reduce the problem that any parcel, particularly a small one, that neighbours a parcel with high woodland
Chapter 6: Attribute measurement within parcels and grown regions

PPL values will automatically be considered a potential increase in woodland. This means that it would be possible, for example, that an open water parcel representing a reservoir next to dense woodland be considered a potential increase in woodland. In the case where a parcel is created by the segmentation process that contains partly open water and partly woodland, that parcel will not be removed by this filter process, and represents a weakness of being constrained by the fixed parcels. By applying this constraint, however, occurrences of parcels that are purely water next to large woodland parcels being considered woodland are minimised, although they are not eradicated due to the nature of the data.

A further filter was applied to ensure that parcels are not duplicated in the calculation and the potential increase of woodland was calculated for the two different definitions of woodland being reviewed. Again the same processes were also carried out with respect to urban cover types, using the three thresholds identified as definitions of urban land cover.

6.3.4 Relationship between areas identified as potential increase in woodland and areas under management by the National Forest Company

The National Forest Company use a variety of methods to encourage planting of woodland areas within the Forest. These methods include the use of various grant schemes, land acquisition, and partnerships with private companies and local authorities. All areas of planting, however it was brought about, will be referred to here as management areas for simplicity. Planting has been going on in such management areas since 1995 (National Forest Co., 2004), so a considerable amount of it has occurred since the imagery used in the production of LCM2000 was created. The areas already planted at the time the images were taken would have been covered by immature trees and so were unlikely to have been correctly
classified as woodland within LCM2000. As such the management areas represent a further increase in woodland beyond the areas identified by the methodology outlined above. As such a comparison between the areas identified as a potential increase and the management areas was carried out.

Creating a union of the potential increase datasets at the two different canopy thresholds, and the management areas performed the comparison. From this new dataset it is possible to identify areas of overlap between them.

6.3.5 Multiple classes methods

A further investigation was carried out in order to assess the occurrence of parcels that are identified as a potential increase for both urban and woodland cover types. This was done by taking the datasets that identified the potential increases for both cover types, using both individual parcel attributes and merged neighbours, and performing an intersection. In the same way it would also be of interest to analyse parcels that were already categorised as one of the categories of interest and were identified as a potential increase in the other. It was decided here to analyse those identified as increases in both categories in order to assess the potential impact of the region growing methodology (see Section 2.1.2.4 for further discussion on multiple classes).

A further method to locate urban woodland was investigated looking at the idea that small parcels that represented the equivalent of residential plots could be identified. These parcels would need to contain some element of a residential class in the per pixel data (17.1.1 or 17.2.1 see Appendix 1), even if this was a small proportion, in order to show they are at least partly residential. They would also need to meet some threshold level of woodland criteria.
This is therefore an alternative method of using the metadata to identify an urban, or more specifically a residential area. Such parcels can then be reviewed to assess if it can be classified as urban-woodland. Characteristics other than the meeting of a threshold are therefore being used to identify developed areas.

This work looking at multiple classes attempts to review the metadata in ways to address issues of semantic confusion between land cover types, specifically between residential and urban types in the latter example (see Section 2.1.2.4). The broader region growing methods described above address the spatial confusion of classifying land cover types within the framework of the parcels created for LCM2000.
Chapter 6: Per Pixel List support for classification within parcels and grown regions

6.4 Results

In reviewing support for and conflict with parcel classification provided by the PPL there are a number of different factors that must be considered. Firstly, what are the different measures of a land cover type available to us, and what is the effect of using these different measures? Secondly, to what extent do using these measures create discord between them and the parcel classification if they are considered using a purely Boolean classification? (see Fisher, 1999 and section 6.8.2 for further discussion). These questions are addressed by initially reviewing the object level data for each parcel, assessing the extent of each reviewed land cover type under such measures. Following this the areas that generate discord between the parcel classification and the PPL under a Boolean system are located. The emphasis is then placed on locating parcels that represent a potential increase in the target land cover type over that reported in the parcel classification. Locating parcels that are erroneously reported as the target land cover type, thus representing a potential decrease in area, is easily achieved using this methodology, but will not be developed here, for clarity. These stages are reported for woodland and urban land cover in turn.

Areas identified as a potential increase by the region growing technique are reported, leading to a measure of the total potential increase in area from the region growing method as well as from the metadata of individual parcels. Once the parcels representing the total potential increase are identified they are compared with the management areas of the National Forest Company. In the final part of this section parcels that are identified as a potential increase in both the woodland and urban analyses are reported as areas that meet both criteria and so can be classified as urban woodland and so viewing these areas in terms of mixed classes rather than a Boolean classification.
6.4.1 Woodland land cover extent

The first step in this review of support and contradiction between the pixel and parcel classifications, in the case of woodland, is to identify the areas of woodland, based on the different classifications carried out in the production of LCM2000 and these are shown in the following figures. Figure 6.3 shows the woodland extent based on the parcel classification, in other words the extent shown in the purchased database. Extents based on the PPL, being whether or not the level of woodland in the PPL meets the criteria being used, are shown in Figure 6.4. By way of comparison, Figure 6.5 shows the woodland extent given in the 1:250,000 mapping published by Bartholomew, a much less detailed dataset than the LCM2000. Table 6.4 gives a comparison of the different areal extents using these different definitions of woodland.

Figure 6.3: The extent of woodland within the National Forest as classified within the LCM2000
Figure 6.4: Parcels in which the PPL woodland values meet the threshold values of a) 50% and b) 20%
Chapter 6: Per Pixel List support for classification within parcels and grown regions

Figure 6.5: Woodland extent as recorded by the 1:200,000 mapping published by Bartholomew

Table 6.4: Extent of woodland land cover under different classification criteria

<table>
<thead>
<tr>
<th>Basis of classification as target land cover type</th>
<th>Area (hectares)</th>
<th>Percentage of the National Forest</th>
<th>Number of parcels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Woodland in LCM2000</td>
<td>4,385.54</td>
<td>8.70%</td>
<td>1,451</td>
</tr>
<tr>
<td>Parcels with PPL &gt; 20%</td>
<td>5,813.19</td>
<td>11.53%</td>
<td>2,053</td>
</tr>
<tr>
<td>Parcels with PPL &gt; 50%</td>
<td>3,347.92</td>
<td>6.64%</td>
<td>1,013</td>
</tr>
</tbody>
</table>

6.4.2 Discord between woodland classification and the per pixel metadata

Discord can occur between the parcel classification and the PPL in two ways when reviewing a particular land cover type. When the parcel is classified as the land cover in question the PPL may not support that classification, in that it does not meet the pre-determined threshold. Alternatively, the PPL may meet the threshold, even though the parcel is not classified as that land cover type. The areas where such discord occurs are shown in Figures 6.6 and 6.7 and described in Table 6.5.
Figure 6.6: Unsupported woodland parcels classified as woodland by the per parcel classification but that have PPL woodland values less than the threshold values of a) 20% and b) 50%

The area shown in Figure 6.6 are those parcels that are classified as woodland in the LCM2000, but that do not meet the woodland criteria within the metadata. Figure 6.6a shows those that do not meet the 20% threshold, which cover 135.20ha. Figure 6.6b shows those that do not meet the 50% threshold, covering 1,170.81ha (see also Table 6.5). These
Chapter 6: Per Pixel List support for classification within parcels and grown regions

represent 3.08% and 26.70% respectively of the 4,385.54ha area classified as woodland by the parcel classification.

Figure 6.7: New woodland parcels that were not classified as woodland by the per parcel classification but that have PPL woodland values greater than the threshold values of a) 20% and b) 50%
Table 6.5: Areas of parcels that have a contradiction between the LCM2000 classification and the Per Pixel List (PPL) relating to classification of woodland land cover

<table>
<thead>
<tr>
<th>Parcels with inconsistent PPL to classification</th>
<th>Area (hectares)</th>
<th>Percentage of the National Forest</th>
<th>Number of parcels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Woodland in LCM2000 and PPL &lt; 20%</td>
<td>135.20</td>
<td>0.27%</td>
<td>77</td>
</tr>
<tr>
<td>Woodland in LCM2000 and PPL &lt; 50%</td>
<td>1,170.81</td>
<td>2.32%</td>
<td>515</td>
</tr>
<tr>
<td>Non Woodland in LCM2000 and PPL &gt; 20%</td>
<td>1,462.04</td>
<td>2.90%</td>
<td>645</td>
</tr>
<tr>
<td>Non Woodland in LCM2000 and PPL &gt; 50%</td>
<td>120.46</td>
<td>0.24%</td>
<td>67</td>
</tr>
</tbody>
</table>

Table 6.6: Per pixel list values for all land cover types from parcels classified as woodland but having woodland per pixel list values that do not meet the threshold criteria of 20% and 50%.

<table>
<thead>
<tr>
<th>Broad Habitat</th>
<th>Deciduous</th>
<th>Coniferous</th>
<th>Arable</th>
<th>Grass Improved</th>
<th>Grass Rough</th>
<th>Grass Acid</th>
<th>Bracken</th>
<th>Dwarf shrub heath</th>
<th>Fen Marsh Swamp</th>
<th>Bog</th>
<th>Water Inland</th>
<th>Inland bare ground</th>
<th>Urban</th>
<th>Less than 20% n = 77</th>
<th>Less than 50% n = 515</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than 20%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Average</td>
<td>Maximum</td>
</tr>
<tr>
<td>Average</td>
<td>10</td>
<td>2</td>
<td>33</td>
<td>13</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>10</td>
<td>50</td>
</tr>
<tr>
<td>Maximum</td>
<td>19</td>
<td>17</td>
<td>89</td>
<td>81</td>
<td>45</td>
<td>11</td>
<td>33</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>14</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Less than 50%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Less than 50% n = 515</td>
<td>Less than 50% n = 515</td>
</tr>
<tr>
<td>Average</td>
<td>28</td>
<td>4</td>
<td>26</td>
<td>10</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>Maximum</td>
<td>49</td>
<td>46</td>
<td>89</td>
<td>81</td>
<td>45</td>
<td>35</td>
<td>38</td>
<td>20</td>
<td>0</td>
<td>71</td>
<td>20</td>
<td>20</td>
<td>62</td>
<td>62</td>
<td>62</td>
</tr>
</tbody>
</table>

The areas shown in Figure 6.7 are the parcels that are not classified as woodland in the LCM2000, but that do meet the woodland criteria within the metadata. Figure 6.7a shows those that meet the 20% threshold and these parcels cover 1,462.04ha. Figure 6.7b shows those that meet the 50% threshold, covering 120.46ha. These represent a potential increase of 33.34% and 2.75% respectively from the 4,385.54ha area classified as woodland by the parcel classification.
Chapter 6: Per Pixel List support for classification within parcels and grown regions

Table 6.6 shows the Broad Habitats that are supported by the per pixel list metadata where it does not support the presence of woodland, even though the parcel is classified as woodland. The pattern shown by the averages at the two thresholds is very similar with the exception of deciduous woodland. This pattern is not particularly surprising as it shows a considerable increase in the support for woodland as the threshold increases. Simply showing that in parcels classified as woodland many have per pixel list values of higher than 20%, even though they do not meet the 50% threshold. Beside the level of per pixel list support for woodland categories the support is mostly for arable land cover and to a lesser extent improved grass and urban. This implies that there is considerable confusion between these cover types, or that these parcels are highly heterogeneous, containing mixtures of these cover types.

6.4.3 Urban land cover extent

For the analysis of urban areas, again the extent of the land cover type as given in the LCM2000 is presented first. The parcel classification gives the extent of urban as that in Figure 6.8 and the PPL gives the extent shown in Figure 6.9 when compared against the three criteria of 35%, 55% and 75% urban cover. These extents are all then compared in Table 6.7. The areas under the three different extents derived from the PPL are all less than that given by the parcel classification, unlike the woodland example, in which the least stringent criteria yields a greater area than the parcel classification.
Chapter 6: Per Pixel List support for classification within parcels and grown regions

Figure 6.8: The extent of urban areas within the National Forest as classified within the LCM2000

Figure 6.9: Parcels in which the PPL urban values meet the threshold values of a) 75%, b) 55% and c) 35%
Figure 6.9 (cont.): Parcels in which the PPL urban values meet the threshold values of a) 75%, b) 55% and c) 35%
Chapter 6: Per Pixel List support for classification within parcels and grown regions

Table 6.7: Extent of urban land cover under different classification criteria

<table>
<thead>
<tr>
<th>Basis of classification as target land cover type</th>
<th>Area (hectares)</th>
<th>Percentage of the National Forest</th>
<th>Number of parcels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urban in LCM2000</td>
<td>7,977.04</td>
<td>15.82%</td>
<td>2,607</td>
</tr>
<tr>
<td>Parcels with PPL &gt; 35%</td>
<td>7,349.33</td>
<td>14.57%</td>
<td>2,337</td>
</tr>
<tr>
<td>Parcels with PPL &gt; 55%</td>
<td>5,457.83</td>
<td>10.82%</td>
<td>1,522</td>
</tr>
<tr>
<td>Parcels with PPL &gt; 75%</td>
<td>3,424.24</td>
<td>6.79%</td>
<td>839</td>
</tr>
</tbody>
</table>

It is clear from Figure 6.9 that the main roads, particularly the M1, which curves through the eastern end of the Forest (see Figure 1.3), are picked out quite clearly when a threshold of just 35% of urban categories in the PPL is used. This shows that the algorithm used to create the parcels must have picked out fairly linear parcels along the main roads, and that many of them contain at least some urban pixels. When the threshold is increased to 75% the locations of the A roads are not distinguishable and the motorway can only be seen along parts of its route. So the per-pixel classification appears to be picking up on the road surface on large roads that are probably wider than the 25m pixels, but also picking up on surrounding cover types such as verges and possibly grassed central reservations. This does seem to be supported by the fact that the areas of the M1 that exceed the 75% urban threshold are sections of the road that are at major junctions, where there is a greater area of road surface, and also sections where there are large cuttings, so the road is bounded by rock faces.

6.4.4 Discord between urban classification and the per pixel metadata

Parcels with conflicting information between the parcel and pixel classifications are reviewed, just as in the case of woodland. Those classified as urban that do not meet the 35%, 55% and 75% thresholds for urban (Figure 6.10 and Table 6.8) represent 20.98%, 37.10% and 59.62% respectively of the area classified as urban by the parcel classification. These figures are considerably larger than the equivalent statistics in the woodland analysis. Those parcels not classified as urban but that meet the same urban PPL thresholds (Figure 6.11 and Table 6.8)
represent 12.21%, 5.30% and 2.10% of the area classified as urban respectively. From this
evidence it would appear that the urban cover type is more likely to be overestimated than
underestimated by the parcel classification.

Figure 6.10: Unsupported urban parcels classified as urban by the per parcel classification but that
have PPL urban values less than the threshold values of a) 35%, b) 55% and c) 75%
Figure 6.11: New urban parcels that were not classified as urban by the per parcel classification but that have PPL urban values greater than the threshold values of a) 35%, b) 55% and c) 75%
Chapter 6: Per Pixel List support for classification within parcels and grown regions

Table 6.8: Areas of parcels that have a contradiction between the LCM2000 classification and the Per Pixel List (PPL) relating to classification of urban land cover

<table>
<thead>
<tr>
<th>Parcels with inconsistent PPL to classification</th>
<th>Area (hectares)</th>
<th>Percentage of the National Forest</th>
<th>Number of parcels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urban in LCM2000 and PPL &lt; 35%</td>
<td>1,673.81</td>
<td>3.32%</td>
<td>760</td>
</tr>
<tr>
<td>Urban in LCM2000 and PPL &lt; 55%</td>
<td>2,959.20</td>
<td>5.87%</td>
<td>1,278</td>
</tr>
<tr>
<td>Urban in LCM2000 and PPL &lt; 75%</td>
<td>4,755.53</td>
<td>9.43%</td>
<td>1,850</td>
</tr>
<tr>
<td>Non Urban in LCM2000 and PPL &gt; 35%</td>
<td>974.23</td>
<td>1.93%</td>
<td>463</td>
</tr>
<tr>
<td>Non Urban in LCM2000 and PPL &gt; 55%</td>
<td>422.66</td>
<td>0.84%</td>
<td>184</td>
</tr>
<tr>
<td>Non Urban in LCM2000 and PPL &gt; 75%</td>
<td>167.57</td>
<td>0.33%</td>
<td>72</td>
</tr>
</tbody>
</table>

Table 6.9: Per pixel list values from parcels classified as urban but having per pixel list values that do not meet the threshold criteria.

<table>
<thead>
<tr>
<th>Broad Habitat</th>
<th>Less than 35%</th>
<th>Less than 55%</th>
<th>Less than 75%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average</td>
<td>Maximum</td>
<td>Average</td>
</tr>
<tr>
<td>Deciduous</td>
<td>4 1 43 5 6</td>
<td>63 45 101 64 56 36</td>
<td>3 1 29 4 3 0</td>
</tr>
<tr>
<td>Coniferous</td>
<td>1 1 10 0 2 0</td>
<td>42 46 26 0 60 82</td>
<td>1 1 10 0 2 0</td>
</tr>
<tr>
<td>Arable</td>
<td>5 1 0 0 0 0</td>
<td>26 0 60 82 74 0</td>
<td>5 1 0 0 0 0 0</td>
</tr>
<tr>
<td>Grass Improved</td>
<td>6 0 0 0 0 0</td>
<td>36 46 42 26 0 60</td>
<td>6 0 0 0 0 0 0</td>
</tr>
<tr>
<td>Grass Rough</td>
<td>0 1 1 0 0 0</td>
<td>26 0 60 82 74 0</td>
<td>0 1 1 0 0 0 0</td>
</tr>
<tr>
<td>Grass Acid</td>
<td>1 1 0 0 0 0</td>
<td>26 0 60 82 74 0</td>
<td>1 1 0 0 0 0 0</td>
</tr>
<tr>
<td>Bracken</td>
<td>0 0 0 0 0 0</td>
<td>26 0 60 82 74 0</td>
<td>0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>Dwarf shrub</td>
<td>0 0 0 0 0 0</td>
<td>26 0 60 82 74 0</td>
<td>0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>Fen Marsh Swamp</td>
<td>2 0 0 0 0 0</td>
<td>26 0 60 82 74 0</td>
<td>2 0 0 0 0 0 0</td>
</tr>
<tr>
<td>Bog</td>
<td>0 0 0 0 0 0</td>
<td>26 0 60 82 74 0</td>
<td>0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>Water Inland</td>
<td>0 0 0 0 0 0</td>
<td>26 0 60 82 74 0</td>
<td>0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>Inland bare ground</td>
<td>0 0 0 0 0 0</td>
<td>26 0 60 82 74 0</td>
<td>0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>Urban</td>
<td>2 1 18 0 0 0</td>
<td>60 82 74 0 0 0</td>
<td>2 1 18 0 0 0 0</td>
</tr>
</tbody>
</table>

The broad habitats supported in the PPL of parcels classified as urban, but having little support for that classification in the per pixel list, are shown in Table 6.9. A very similar
pattern to that displayed in the woodland analysis (Table 6.6) emerges, with a very static structure to the data, except for the average values for urban increasing as the threshold increases. As in the woodland example this is to be expected, but in this case there is really only one other land cover type being supported by the PPL in these parcels, and that is arable.

6.5 Region Growing Results

In order to break down the fixed spatial unit of the parcel the region growing methodology was carried out as outlined in Section 6.3.3. The results of this analysis are reported, giving the woodland land cover type first, then the urban. Within each of these, the parcels representing potential increase in the cover type when merged are recorded and then the total increase, which includes the individual parcels not classified as the target type, but with support for the type in the PPL (Figures 6.7 and 6.11) as well as those identified in the region growing.

6.5.1 Woodland

The region growing method using merged neighbouring polygons was used to identify parcels that represent potential increase in the woodland area within the National Forest as reported by the parcel classification of LCM2000. The output of this analysis is shown in Figure 6.12 for the two thresholds used in the woodland example. In order to get a complete picture of the potential increase based on the support from the PPL the parcels identified by the region growing technique must be added to the parcels identified as a potential increase in Figure 6.7 i.e. those parcels with a PPL that meets the criteria without merging. By performing a union of these datasets a final total potential increase is derived, which is only slightly different from the output shown in Figure 6.12, which is illustrated by Table 6.10. This shows that almost all parcels that meet the criteria based solely on their own PPL are identified as
potential increase by the region growing method. Only 145.8ha are not identified using the 20% threshold. These polygons are surrounded by polygons with very low woodland PPL values in order for them not to be identified by the region growing method. As the result of the total potential increase is very similar to Figure 6.12, no further illustration has been included.

An example is shown of an area identified as a potential increase by the region growing technique in Figure 6.13. It shows an area of quite young planted trees, both deciduous and coniferous, with older trees beyond. The conifers would have been only a year or two old when the imagery was taken that was used to create LCM2000, so the spectral signature would have been more like that of shorter vegetation, such as the grassland that it was actually categorised as. Some of the polygon however, was categorised as woodland in the PPL and so the technique was able to pick it out as potential woodland.

Table 6.10: Potential increase in woodland areas from region growing analysis of PPL

<table>
<thead>
<tr>
<th></th>
<th>20% threshold</th>
<th>50% threshold</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Increase</td>
<td>Total</td>
<td>Difference</td>
<td>Increase</td>
<td>Total</td>
</tr>
<tr>
<td></td>
<td>when merged</td>
<td>increase</td>
<td></td>
<td>when merged</td>
<td>increase</td>
</tr>
<tr>
<td>No. of parcels</td>
<td>1,408</td>
<td>1,514</td>
<td>106</td>
<td>574</td>
<td>585</td>
</tr>
<tr>
<td>Hectares</td>
<td>3,752.67</td>
<td>3,898.47</td>
<td>145.80</td>
<td>1,183.11</td>
<td>1,199.98</td>
</tr>
</tbody>
</table>
Figure 6.12: Parcels that are not classified as woodland in the LCM2000 but are potentially woodland if merged with their neighbour as the merged parcel has PPL woodland values that meet the threshold values of a) 20% and b) 50%
Figure 6.13: Example of area identified as potential increase in woodland land cover using the threshold of 20%. The foreground area was classified by the parcel classification as “Grass improved intensive” (5.1.1, ID: SK040742r1). Photograph taken from E450631 N310144, direction 35 degrees, April 2004.

6.5.2 Urban

As with the woodland example the area of potential increase from the region growing method was ascertained (Figure 6.14). A union operation was then performed with the potential increase from the support within individual parcels (Figure 6.11), giving a total potential increase. This final dataset is compared with the potential increase from the region growing alone (Table 6.11), with the difference column showing the parcels with enough support individually, but not when merged with neighbours. Again as the distribution of the total potential increase is very similar to Figure 6.14 no further illustration has been included.

The example of a parcel identified as a potential increase in urban area (Figure 6.15) shows the car park of a local sports ground. While this is clearly an urban land cover, with a large proportion of impervious surface, and is on the edge of a built up area, it is adjacent to a large sports field on one side and the covered reservoir that can be seen in the background of the
picture on another. It would appear then that this parcel is an example of an area on the fringe of urban land cover, where classification uncertainty is higher because of semantic uncertainty relating to how urban land is defined, as well as technical problems with classifying mixed areas (see Section 6.8.1 for further discussion).

Figure 6.14: Parcels that are not classified as urban in the LCM2000 but are potentially urban if merged with their neighbour, as the merged parcel has PPL urban values that meet the threshold values of  

a) 35%, b) 55% and c) 75%
Chapter 6: Per Pixel List support for classification within parcels and grown regions

Table 6.11: Potential increase in urban areas from region growing analysis of PPL

<table>
<thead>
<tr>
<th></th>
<th>35% threshold</th>
<th></th>
<th></th>
<th>55% threshold</th>
<th></th>
<th></th>
<th>75% threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Increase when merged</td>
<td>Total increase</td>
<td>Difference</td>
<td>Increase when merged</td>
<td>Total increase</td>
<td>Difference</td>
<td>Increase when merged</td>
</tr>
<tr>
<td>No. of parcels</td>
<td>1,333</td>
<td>1,411</td>
<td>78</td>
<td>691</td>
<td>735</td>
<td>44</td>
<td>224</td>
</tr>
<tr>
<td>Ha</td>
<td>3,094.64</td>
<td>3,211.03</td>
<td>116.39</td>
<td>1,500.13</td>
<td>1,571.38</td>
<td>71.25</td>
<td>448.93</td>
</tr>
</tbody>
</table>

Figure 6.15: Example of area identified as potential increase in urban land cover. The foreground area was classified by the parcel classification as “Grass calcareous rough” (7.1.2, ID: SK124910r1). Photograph taken from E444405 N313505, direction 100 degrees.
Chapter 6: Per Pixel List support for classification within parcels and grown regions

6.6 Relationship between areas identified as potential increase in woodland and areas under management by the National Forest Company

The datasets created by performing a union of the potential increase areas and the woodland management areas are shown in Figures 6.16 and 6.17. This provides a useful analysis for two reasons. Firstly, areas outside the management areas may be of interest to the National Forest Company as areas of existing woodland that they may not have currently mapped, which is an important issue as the company has targets to achieve in increasing the proportion of the area under woodland. Secondly, the management areas will over time become woodland, so will represent an increase in woodland over forthcoming years and this is already known because of the planned management of these areas. Table 6.12 shows that the majority of areas identified as potential increase, when using either threshold, are outside the management areas, indicating that the identified areas could make a significant impact on the National Forest Companies targets. It can also be seen from Table 6.12 what the potential increase will be as the management areas become woodland.
Figure 6.16:
Relationship between areas identified as a potential increase in woodland using a 50% canopy threshold criteria and the areas under management by the National Forest Company. Also shown are the areas originally classified as woodland by the LCM2000 (per-parcel) and the relationship with the management areas.
Figure 6.17: Relationship between areas identified as a potential increase in woodland using a 20% canopy threshold criteria and the areas under management by the National Forest Company. Also shown are the areas originally classified as woodland by the LCM2000 (per-parcel) and the relationship with the management areas.
Table 6.12: Comparison of areas identified as a potential increase in woodland with those areas under management by, or licence agreement with, the National Forest Company

<table>
<thead>
<tr>
<th>Management areas not identified as a potential increase</th>
<th>Areas of potential increase outside management areas</th>
<th>Areas of potential increase within management areas (intersection)</th>
<th>Total area (excluding intersection)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area in hectares</td>
<td>4,432.62</td>
<td>1,077.22</td>
<td>122.78</td>
</tr>
<tr>
<td>Intersection as a percentage</td>
<td>2.77</td>
<td>10.23</td>
<td></td>
</tr>
</tbody>
</table>

Potential increase identified using 50% threshold criteria

| Area in hectares                                        | 4,124.27                                         | 3,468.77                                                      | 431.13                           | 7,593.04                          |
| Intersection as a percentage                            | 9.46                                             | 11.06                                                        |                                  | 5.68                              |

6.7 Multiple classes results

Only small areas were located that meet thresholds for both woodland and urban classes (Table 6.13). They are so small that they do not show clearly on a map of the whole forest, so one is not reproduced here. The single parcel that meets the most stringent criteria for both classes is shown in relation to OS 1:50,000 mapping in Figure 6.18, and it clearly occupies a boundary area between a housing development and a small deciduous woodland. Individually this parcel has PPL values of just 31% suburban/rural developed, 16% broadleaved woodland (split evenly between deciduous and mixed), 31% bare ground – semi-natural and 8% arable unknown. So if considered on its own this parcel would not reach the lowest threshold for either urban or woodland (being 35% and 20% respectively).

Figure 6.19 shows those parcels that meet the two lowest thresholds for urban and woodland used in this analysis (pink) and the parcels classified as urban by the per parcel classification around the area of Coalville. A clear pattern is displayed showing the urban woodland parcels around the periphery of the urban areas.
Table 6.13: Statistics of the parcels that meet criteria for both urban and woodland at different threshold levels

<table>
<thead>
<tr>
<th>Urban threshold</th>
<th>Woodland threshold</th>
<th>Number of parcels</th>
<th>Total area (ha)</th>
<th>Mean parcel size (ha)</th>
</tr>
</thead>
<tbody>
<tr>
<td>75%</td>
<td>50%</td>
<td>1</td>
<td>0.81</td>
<td>0.81</td>
</tr>
<tr>
<td>55%</td>
<td>50%</td>
<td>7</td>
<td>13.30</td>
<td>1.90</td>
</tr>
<tr>
<td>55%</td>
<td>20%</td>
<td>61</td>
<td>135.85</td>
<td>2.23</td>
</tr>
<tr>
<td>35%</td>
<td>20%</td>
<td>177</td>
<td>397.79</td>
<td>2.25</td>
</tr>
</tbody>
</table>

Figure 6.18: Location of the single parcel (SK248360r1 shown in pink near the centre of the map) that meets the 75% urban threshold and the 50% woodland threshold when merged with neighbours. Shown over OS 1:50,000 map data.
Figure 6.19: Location of the parcels that meet the 35% urban threshold and the 20% woodland threshold, either individually or when merged with neighbours (shown in pink). Parcels designated as urban by the parcel classification are shown in blue. Red circle shows parcel detailed in Figure 6.20. Backdrop is OS 1:50,000 map data.

Figure 6.20: Parcel identified as potential increase in both urban and woodland. A red circle locates the parcel in Figure 6.19. a) parcel in the distance, containing the tall trees to the right of the large building; b) the southern end of the parcel.

An example of a parcel identified as urban woodland is shown in Figure 6.20. The large building in Figure 6.20a is the leisure centre marked in Figure 6.19, and the parcel takes in a tree lined stream as well as the houses and road around it. Figure 6.20b shows trees at the southern end of the parcel along the bank of the stream where a small car park abuts the
Both images show how well established trees occupy the riparian zone but are closely related to the urban surroundings of residential and leisure land use. It is partly a question of scale or resolution of the map as to whether the area of trees should be depicted as a woodland parcel or whether it be simply a part of the urban area. It is also a question of definition (Taylor, 1982 cited in Fisher, 1999).

The second analysis outlined in section 6.4.5, which selects any parcels with residential classes in the PPL to identify possible areas of urban woodland was reviewed by overlaying the output over Ordnance Survey 1:50,000 maps and although not analysed systematically many of the parcels identified appeared to be relating to roads rather than settlement areas. On the basis of this the analysis was not taken further.
6.8 Discussion

6.8.1 Land cover extent by different measures

The extent of urban and woodland land cover as measured by the PPL show logical patterns in both cases, as we see considerable increases in extent as the threshold is lowered (Tables 6.4 and 6.7). With both land cover types the area covered at a number of sampled threshold values follows a linear trend and the areas covered by the parcel classifications would equate to using a 36% and 32% threshold for woodland and urban respectively. This would seem to imply that the higher thresholds used may be too stringent, if the parcel classification is assumed to be a reasonably accurate representation, but the woodland extent based on a 50% threshold interestingly picks out many of the areas that are given by the Bartholomew 1:200,000 dataset (Figure 6.5). The more general dataset will only depict larger areas of woodland, so a high threshold appears to still give a reasonable response when considered on a smaller scale. This also underlines the observation of decreasing uncertainty in larger contiguous areas, which supports findings in Chapter 5.

The concept that uncertainty decreases in continuous areas of a land cover is further supported by the urban results (Figure 6.9), which show that as the threshold is increased to 75% many of the smaller areas disappear, as well as areas around the periphery of towns, leaving stronger responses within the cores of more significant towns, which is where we have noted there is less uncertainty about the definition of what is urban (Zhang and Goodchild, 2002, p90). This, in itself, would seem to be useful information about the presence of significant built up areas such as small towns; as such areas persist when high thresholds of the urban PPL values are examined. A further observation about the effect of increasing the threshold in the urban analysis is made in section 6.4.3, relating to the detection of the main roads, which are only clearly depicted at the lowest threshold. At the higher thresholds they do not
Chapter 6: Per Pixel List support for classification within parcels and grown regions

Persist, giving further evidence that parcels with very high urban PPL values are mainly indicating cores of built up urban areas. It also shows us that in order to obtain a realistic image of the urban land cover a fairly low threshold is required, as Table 6.7 shows, even a threshold of just 35% yields an area identified as urban that is smaller than the area identified under the per-parcel classification. This is likely to be related to the fact that urban land cover tends, by its very nature, to be very heterogeneous, particularly in suburban areas where buildings are less dense and there is more vegetation (Hung, 2002, Small, 2001). Heterogeneity of land cover leads to the problem of mixed pixels (mixels), which occur when a single pixel relates to an area with more than one land cover type. Mixels lead to inaccuracy in classification and are commonly a problem in the urban environment (Forster, 1983; Mather, 1987, Zhang and Goodchild, 2002). As a consequence training a classifier to identify a category of “urban” and particularly “suburban”, is extremely difficult due to the wide range of possible spectral responses in such areas.

6.8.2 Discord between the parcel classification and the per pixel metadata

Discord occurs where “one object or individual is clearly defined but is shown to be a member of two or more different classes under differing schemes or interpretations of the evidence” (Fisher, 1999; see Figure 2.1). In the LCM2000 we have two interpretations of the data, the per-parcel classification and the per-pixel classification. The per-parcel classification follows a Boolean model and where we consider the per-pixel scheme as meeting a threshold or not we are viewing that under a Boolean model also. Therefore where these two schemes do not agree, we are left with discord. While it is not suggested here that the reaching of a threshold is the only way the PPL data can, or should, be viewed, it does however provide an interesting analysis by reviewing the level of discord generated.
Chapter 6: Per Pixel List support for classification within parcels and grown regions

Over 26% of the area categorised as woodland in the parcel classification is not supported by PPL values of greater than 50% (Section 6.4.2). This appears to be a very high figure, leading to large areas of discord. Such a result could be indicating that large areas classified as woodland are not densely wooded, or it could be suggesting large uncertainty in the classification of woodland. The corresponding percentages for urban parcels are similarly high (Section 6.4.4). This is perhaps not surprising because of the heterogeneity of urban areas, but it does raise considerable questions about the reliability of urban categories, particularly outside of core areas of towns, in more rural settings or in urban fringes.

The level of discord between the two classifications is highest between arable classes and the two target land cover types (Tables 6.6 and 6.9). With urban parcels it is really only the arable class that causes substantial levels of discord. In parcels classified as urban, but with little supporting evidence in the PPL, there are generally low values for woodland (Table 6.9). In the opposite case, however, there are slightly higher PPL values for urban within woodland parcels (Table 6.6), which also have similar values for improved grassland. In such parcels, where discord exists, the confusion is therefore primarily between woodland and arable, improved grass or urban, and between urban and arable. This is consistent with the findings from look up tables (LUT's) drawn up by experts asked to identify where confusion is likely or possible between land cover types, based on technical issues (such as reflectance) (Comber pers. comm.). The table identifies that there is likely to be confusion between broadleaved woodland and both tilled land and suburban areas, two of the largest confusions identified in Table 6.6. It also suggests that there would be possible confusion between coniferous woodland and tilled land, as well as between both kinds of woodland and grazed turf, rough grass and heath. While these categories are not exactly the same as those used in LCM2000, the reason for this being that the LUT was intended to compare LCMGB with LCM2000, it is still indicative of the areas where technical issues can create confusion in classification.
These results do very clearly pick out known issues that LCM2000 has in distinguishing woodland from arable in early season imagery and suburban areas from early crops due to their phenological similarities (Smith pers. comm.).

The strong confusion between urban and arable categories shown in Table 6.9 are less well explained by the expert LUT, as the urban categories are only identified as having possible confusion with bare ground and deciduous trees. The LUT does, however, identify possible confusion between arable and bare ground categories, so it is possible that the pattern shown in Table 6.9 is relating to a complex relationship between the three categories of urban, arable and bare ground.

6.8.3 Region growing

Developing the region growing technique has been a very important part of this analysis for a number of reasons. The spatial unit of the parcel within the LCM2000 is fixed but potentially meaningless due to the method of its construction. That is not to say that they are all meaningless as recognisable shapes can be seen within the parcel structure in many places, but they relate purely to the analysis of spectral response, not of recognisable geographical features. Given this situation, the ability to break down the boundaries of these parcels prevents us from being constrained by their fixed nature. It also gives the opportunity to reuse the data contained in the PPL in a potentially more informative way.

Using the region growing technique to locate parcels that are a potential increase in woodland gives substantial areas that meet the chosen criteria. It is important that when looking for potential increase the parcels that meet the criteria individually, without being merged with neighbours are included, although most already are by the nature of the process. Once all the
relevant parcels are taken into consideration the increase with a 20% threshold is 3,898ha, which represents an 88% increase from the 4,406ha identified as woodland by the parcel classification. Even at the 50% threshold the potential increase is still over 27%. The equivalent percentage increases for the urban thresholds are 40%, 19% and 6%, so the technique is still locating areas using the very high threshold of 75% in a class like urban, which by its very nature tends to be highly heterogeneous.

Examples of parcels located by the region growing technique are given in Figures 6.13 and 6.15. The woodland example in Figure 6.13 shows a parcel containing young trees in an area planted by the National Forest Company. This planting scheme is called Bailey-Sim Wood and the planting was carried out in 1997. Metadata in the LCM2000 tells us that the parcel was classified from imagery taken in 1998, just one year after the trees had been planted. It is unsurprising then that the classification did not generate a response of woodland. The woodland value from the PPL is just 2%, so it would also not be identified by any analysis of the metadata for the individual parcel. Only by breaking down the fixed nature of the parcel is it picked out as potentially a wooded area.

There have been large amounts of tree planting carried out over the last 25 years or more for conservation, improving recreational areas and to make use of unused land (Petit et al, 2004). The region growing method has picked up areas of recent planting (Figure 6.13), so could give a quick indication of changes, which is an important issue as studies of data derived from the Countryside Surveys have shown how dynamic patches of woodland can be over a period as short as 14 years (Petit et al, 2004).

The urban example given in Figure 6.15 is a car park next to sports fields, part of which is classified as urban and part as rough calcareous grassland. As was pointed out in Section
6.5.2, this example is at a boundary between clearly urban land cover, being residential and
car park, and more open types, being unimproved grass on a covered reservoir, trees and a
sports field. The latter could clearly be defined as an urban land use, but in terms of land
cover is distinctly different from that of the car park. As the land parcel created by the
segmentation algorithm includes the car park as well as part of the sports field and the trees
that can be seen in Figure 6.15; it therefore contains mixed land cover and this has apparently
caued problems with the classification, despite it being dominated by the car park. Such
problems are therefore partly caused by the segmentation process creating a parcel that is
mixed, but also by following a Boolean process whereby a single final category must be
allocated. Mather (1999) describes this issue of allocating mixed pixels (in this case mixed
parcels) to one or other candidate land cover class as the major argument for the use of ‘soft’
classification methods, which move away from the strict Boolean model.

6.8.4 Comparison with management areas

Very few of the areas under management agreements with the National Forest Company are
identified as potential increases in woodland (Table 6.12). This is partly because the planting
only started in 1995 and much of the imagery used in the production of the LCM2000 was
taken in 1998, so much of it was carried out after the imagery was taken. It is also likely that
many of the management areas were established in locations that were not wooded.

The implication of this is that the management areas are definitely an increase in woodland
cover (see Chapter 4) and the areas identified in this analysis are also likely increases, so the
total increase could be considerable.
6.8.5 Multiple classes

Fisher (1999) gives a conceptual model of uncertainty in spatial data (see Chapter 2). Within this model one element is called discord, which is where one clearly defined object can be a member of more than one class using different interpretations of the available evidence. Discord is a form of ambiguity, which as Fisher notes is confusion generally caused by differing classification systems. By reviewing a parcel’s possible class based on the PPL as well as the parcel classification, as shown in Figures 6.6, 6.7, 6.10 and 6.11, could lead to a problem of discord. In this case the problem is not one of differing classification systems, as the same one is used in both cases, but one of interpretation of available evidence. The analysis here, while recognising this problem, does not simply accept it and detail areas in which discord occurs. Instead the problem is resolved by recognising the need in these locations to amend the classes used by creating a multiple class, the example here being that of urban woodland. By amending the classes in this way the heterogeneity of these locations is recognised and the description of them therefore improved.

Areas identified as urban woodland are shown in Figure 6.19 and clearly occupy areas around the margins of the urban area. Zhang and Goodchild (2002, p90) cite the example of the gradation from rural to urban areas when discussing the uncertainty surrounding transition from one property to another. They point out that in such cases “there is more likely to be uncertainty regarding the boundaries of their spatial and semantic extents”. It is interesting then that the example given in Figure 6.19 clearly shows how the approach to detect multiple class parcels is generally itemising parcels around the periphery of the obviously urban areas. Other peripheral areas not picked out by this example could well be selected by a similar approach assessing multiple classes of urban with arable or grassland classes.
The land use map of London created by Milne in the 1790's (see Section 2.3; Barber 2005) appears to have followed this sort of approach. It used Ordnance Survey mapping as a base and colour washes were applied to represent land use categories as well as letter codes added to each parcel. While the colours were not mixed, the letter codes do appear to have been. In this way the reader was being informed that more than one class might legitimately describe the parcel, or parts of it. Land use and land cover maps produced in more recent times have not reported mixed classes in this way. The approach seems to have been lost or discarded despite being very rich in information.

6.9 Conclusion

The technique outlined here shows considerable promise in analysing greater complexity in land cover than is possible from a standard land cover map. Data in the PPL can be used to provide an analysis of uncertainty in classification, it can also be re-used to analyse areas beyond the fixed spatial entities of the published parcels. This can identify areas where classification uncertainty may be problematic to a user and also identify areas that could be more accurately described using mixed classes to represent the heterogeneity at the location.

6.10 Further work

The work reported here has produced a method for analysing complexity in classification uncertainty. It would be of considerable use to produce automated algorithms to carry out the region growing aspect of the work, so it could be reproduced quickly for other land cover types. Another point of interest would be to carry out a comparison between the PPL and the V-I-S model, using fieldwork to assess whether the PPL could reliably identify where a parcel falls within the V-I-S continuum.
Chapter 7: Uncertainty metadata in the Land Cover Map 2000 – conclusions and recommendations

7.1 Aims and objectives of the thesis

The specific aim of this thesis was to examine if there is benefit in providing object-based measures of heterogeneity in spatial objects, as support for the reporting of object-based metadata (Section 2.7). In reviewing the problem of how the metadata in LCM2000 can be used there are three constituent parts to the work required. These are using the metadata to identify and describe uncertainty, exploring what new information can be generated from the metadata and, importantly, validating the metadata to ensure that it is informing the user in the way anticipated. Each of these parts has been addressed within the terms of the specific objectives laid out in Section 1.4, showing very positively that elements of the metadata can be used to identify and describe uncertainty, that it can be used to produce new information regarding uncertainty and that it is indeed valid data to use for these purposes (Robinson et al, 2005).

7.1.1 Identification, description and modelling, visualisation and prediction

Section 2.6 pointed out Zhang and Goodchild’s description of how research on uncertainty deals with its identification, description and modelling, visualisation and also with predicting the effects of uncertainty on any analysis carried out, if some knowledge is possessed of uncertainty within the input data (Zhang and Goodchild, 2002, p8). The per-pixel list (PPL) element of the LCM2000 metadata facilitates all of these aspects of research on uncertainty and the objectives of this work follow these lines very closely. The stated objectives of this work were to explore the metadata to see how it can describe uncertainty, to assess how it can
be used to visualise uncertainty, to assess it’s validity and to see what new information can be generated from it. Exploring the metadata can identify uncertainty as well as describe it. The second objective clearly deals with the visualisation aspect. Assessing the validity of the metadata relates less directly to understanding the uncertainty in the map itself and more to realizing how the metadata can aid that understanding. The final objective of generating new information is describing and modelling the uncertainty and this includes information that can be used to predict the effects of the uncertainty.

7.1.1.1 Exploration of the metadata

Identifying uncertainty in data is a complex issue, as described in Chapter 2. Uncertainty can take a variety of forms (see Fisher’s model described in Section 2.1.1); so can be introduced by issues such as ambiguous class definitions or error. Any stage in the production process of a map can therefore be scrutinized in order to identify uncertainty. The main concern of this work however is the classification uncertainty in land cover, assessing where this may occur and how users can assess its impact on their work.

Exploring the PPL for a study area, as was shown in Chapter 3, allows a user to identify whether or not certain broad habitats have greater uncertainty in their classification and as the metadata is object-based this can also be used to assess variation across the area. These techniques can also demonstrate differences between two or more areas by allowing comparison. The publication of two independent classifications also enables users to recognize uncertainty by looking for discord between them.
Chapter 7: Uncertainty metadata in the Land Cover Map 2000 – conclusions

7.1.1.2 Visualisation of the metadata

Visualisation is an important element of geographic information science as it is vital in order to portray spatial information to the user. All aspects of the work in this thesis have used visualisation, but in terms of visualising uncertainty Chapter 3 has shown that elements of the LCM2000 metadata, particularly the PPL, can be used to generate information for the user. Maps can be created of the metadata showing PPL values or the locations of knowledge based corrections, which can be used in conjunction with the map itself. The PPL can also be used to generate new realisations of an area, from which it is possible to create an animation of uncertainty based on data relating to each individual parcel.

7.1.1.3 Validation of LCM2000 metadata

An important step in this work was assessing if the metadata, and particularly the PPL values give a reliable indication of classification uncertainty. This was achieved by carrying out a comparison between the PPL values and the level of cumulative evidence from a series of other datasets (Chapter 4). A strong relationship was demonstrated, which shows that the PPL is giving the user good information regarding this aspect of uncertainty. The per-parcel data from the maximum likelihood classifier was shown to be less reliable in this regard, probably due to the methods used to create the map, and so should not be used in the same way. That is not to say, however, that it is not useful metadata to include at an object level.

7.1.1.4 Generation of new information

New information has been generated from the metadata in nearly all elements of this work. Alternative realisations of the map were created in Chapter 3, which were used to produce an animation, but these could also form the basis of further statistical analysis or to explore error
propagation. Index values were calculated from the PPL values in Chapter 5, which make the metadata easier to use and draw into a single map. They also make it relatively straightforward to analyse the effects of heterogeneity across land cover types and across space. In Chapter 6 completely new spatial units were generated based on combining the metadata of neighbouring parcels. This sort of analysis would not be possible based on dataset level metadata such as would be derived from an error matrix.

7.2 Further work

In several areas of the thesis it has been noted that there are areas of work that would be potentially interesting to pursue further or to approach in different ways. Three of these areas in particular are worthy of note. Firstly the animation described in Chapter 3 could be improved by developing a program that would interact directly with the PPL to produce a continually changing image. While the animation presented here is an adequate example and presents the data in an acceptable way, the alternative approach would allow greater control over the image and have smoother transitions.

The region-growing algorithm presented in Chapter 6 could be taken further by developing a program to automate the process and also extend the algorithm to search out larger groups of parcels that could be merged, looking beyond the first neighbour. This would also allow the user to experiment with different cut off levels and compare the variation in output based on these changes. Finally, it is felt that the work comparing the biological data from the National Biodiversity Network with the indices of dominance (Chapter 5) could be taken further, primarily by increasing the amount of data and by accounting for obvious variation in survey effort.
7.3 Final thoughts

The work outlined in this thesis has thoroughly reviewed elements on the metadata published in the LCM2000, particularly that of the per-pixel list, which is a novel type of metadata regarding data quality. It has been shown to provide a good indication of classification uncertainty and to be a functional attribute in that it can be used in a variety of ways to describe and display uncertainty, to allow comparison between areas and to form the basis of models to further explore the uncertainty in the map. The PPL is therefore extremely useful, even though it is not complete.

The PPL is novel but has been shown here to be a very valuable item of metadata as it is independent of the parcel classification and reported at an object level. Such data quality information being reported at the object level gives far more options to users when it comes to reviewing uncertainty and means that any analysis can be spatially explicit.

One possible improvement to reporting the most common five categories found within each parcel would be to publish both the parcel classification and the pixel classification in full, in a single product. This would give users the opportunity to tailor the sort of work that is in this thesis to the problems and questions that they have. One drawback to this is that users would need a greater level of competency in using software such as GIS, though this is increasingly the case.

This work demonstrates the benefits of object-based metadata. It can be a powerful tool and given the increase in available memory and processing power in computers over recent times, it should be an approach used increasingly in the future. A more difficult task will be to convince people to make use of it.
References


References


216


Hung, Ming-Chih (2002). *Urban land cover analysis from satellite images*. International symposium on future intelligent earth observing satellites, Washington DC.


References


References


Appendix 1

Definitions of broad habitats as used in LCM2000 and list of land cover codes

**Source:** Full LCM2000 dataset description published by the Centre for Ecology and Hydrology
### Broad Habitats (BHs) and their distinction in LCM2000.

<table>
<thead>
<tr>
<th>No.</th>
<th>Broad Habitats</th>
<th>Definition and Distinction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Broad-leaved, mixed and yew woodland</td>
<td>Broad-leaved, in stands &gt; 5 m high with tree-cover &gt; 20%; or scrub &lt; 5 m and yew woodland with cover &gt; 30%. Mixed woodland is included if broad-leaved trees in conifers cover &gt; 20%. Stands ≥ 0.5 ha are mapped as separate blocks.</td>
</tr>
<tr>
<td>2.</td>
<td>Coniferous woodland</td>
<td>Coniferous woodland, semi-natural and plantations, with cover &gt; 20%, and recently felled forestry. Once felled areas are colonised by rough grass, heath or scrub, they take that class.</td>
</tr>
<tr>
<td>3.</td>
<td>Boundaries and linear features</td>
<td>Larger linear features such as shelter belts or motorways; smaller linear features (hedges, walls, smaller roads) are only recorded by the field survey.</td>
</tr>
<tr>
<td>4.</td>
<td>Arable and horticulture</td>
<td>Annual crops, recent leys, freshly ploughed land, rotational setaside, and perennial horticulture crops such as berries and orchards. Once setaside is substantially vegetated with weeds or rough grass, it is included in the Improved grassland Habitat.</td>
</tr>
<tr>
<td>5.</td>
<td>Improved grassland</td>
<td>Improved grasslands in swards dominated by agriculturally 'preferred' species, generally 'improved' by reseeding and/or fertilizer treatment. May be used for agriculture or amenity. Fertile pastures with <em>Juncus effusus</em> are included. Setaside grass is included but, where possible, distinguished at the subclass level; abandoned or little-managed Improved grasslands may be confused with semi-natural swards.</td>
</tr>
<tr>
<td>6.</td>
<td>Neutral grassland</td>
<td>Acid, neutral and calcareous semi-natural swards are generally not reseeded or fertilizer treated; they are dominated by lower productivity grasses, perhaps with many herbs. Grassland management may obscure distinctions from Improved grassland. Neutral, calcareous and acid components are distinguished at subclass level using a soil 'acid sensitivity' map. Pastures with <em>Juncus effusus</em> and with semi-natural spectral-characteristics are included with acid swards.</td>
</tr>
<tr>
<td>7.</td>
<td>Calcareous grassland</td>
<td></td>
</tr>
<tr>
<td>8.</td>
<td>Acid grassland</td>
<td></td>
</tr>
<tr>
<td>9.</td>
<td>Bracken</td>
<td>The bracken Habitat is, at the height of the growing season, dominated by <em>Pteridium aquilinum</em>. Where images pre-date the late growing season, or where stands are dissected, bracken may be missed.</td>
</tr>
<tr>
<td>10.</td>
<td>Dwarf shrub heath</td>
<td>Ericaceous species and gorse forming &gt; 25% of plant cover; open and dense heaths are divided at subclass level. The Habitat includes wet and dry categories but ericaceous vegetation on peat ≥ 0.5 m deep is recorded as 'bog'. In contrast, LCMGB 1990 used a definition based on presence of seasonal standing water.</td>
</tr>
<tr>
<td>11.</td>
<td>Fen, marsh and swamp</td>
<td>Vegetation which is permanently, seasonally or periodically waterlogged. Swamps, fens and flushes are seldom extensive enough to map from satellite images. Rush pastures are more extensive. The category does not include fertile pastures with <em>Juncus effusus</em>.</td>
</tr>
<tr>
<td>12.</td>
<td>Bog</td>
<td>Bogs include ericaceous, herbaceous and mossy vegetation in areas with peat &gt; 0.5 m deep; ericaceous bogs are distinguished at subclass level. Inclusion of Ericaceous bogs contrasts with LCMGB 1990 where bogs were herbaceous or mossy in seasonal standing water.</td>
</tr>
<tr>
<td>13.</td>
<td>Standing open water and canals</td>
<td>Water bodies ≥ 0.5 ha are mapped, but only the wider canals and rivers (&gt;50 m) are shown. LCM2000 does not distinguish standing from flowing water.</td>
</tr>
<tr>
<td>14.</td>
<td>Rivers, streams</td>
<td></td>
</tr>
<tr>
<td>15.</td>
<td>Montane Habitats</td>
<td>Prostrate dwarf heath, sedge and rush, moss heaths and snow bed communities. Limited access during field reconnaissance may limit the accuracy of distinctions.</td>
</tr>
<tr>
<td>16.</td>
<td>Inland rock</td>
<td>Natural and man-made bare ground, including waste tips and quarries.</td>
</tr>
<tr>
<td>17.</td>
<td>Built-up areas and gardens</td>
<td>Urban land, rural development, roads, railways, waste and derelict ground, including vegetated wasteland, gardens and urban trees. In LCM200, all larger areas of vegetation ≥ 0.5 ha are identified as the appropriate cover class. Continuous urban and discontinuous suburban cover are distinguished at subclass level.</td>
</tr>
<tr>
<td>18.</td>
<td>Supra-littoral rock</td>
<td>Supra-littoral Habitats, created by coastal processes of erosion and/or accretion, lie above mean high water spring tides; distinction used a maritime mask. Separation of sediment rock and sediment was at subclass level, through spectral and interactive processing.</td>
</tr>
<tr>
<td>19.</td>
<td>Supra-littoral sediment</td>
<td></td>
</tr>
<tr>
<td>20.</td>
<td>Littoral rock</td>
<td>Littoral Habitats lie below mean high water spring tides in a zone defined by a maritime mask. Rocks and sediments were separated at subclass level by semi-interactive processing. Littoral rocks are generally limited in extent; sediments may be extensive. Saltmarsh is included with Littoral sediments, but as a separate subclass.</td>
</tr>
<tr>
<td>21.</td>
<td>Littoral sediment</td>
<td></td>
</tr>
</tbody>
</table>
Variant classes and codes used in LCM2000

1.1.1 Broad-leaf woodland deciduous
1.1.2 Broad-leaf woodland mixed
1.1.3 Broad-leaf woodland open birch
1.1.4 Broad-leaf woodland scrub
2.1.1 Coniferous woodland
2.1.2 Coniferous woodland felled
2.1.3 Coniferous woodland new plantation
4.1.1 Arable barley
4.1.2 Arable maize
4.1.3 Arable oats
4.1.4 Arable wheat
4.1.5 Arable cereal (spring)
4.1.6 Arable cereal (winter)
4.2.1 Arable arable bare ground
4.2.2 Arable carrots
4.2.3 Arable field beans
4.2.4 Arable horticulture
4.2.5 Arable linseed
4.2.6 Arable potatoes
4.2.7 Arable peas
4.2.8 Arable oilseed rape
4.2.9 Arable sugar beet
4.2.10 Arable unknown
4.2.11 Arable mustard
4.2.12 Arable non-cereal (spring)
4.3.1 Arable orchard
4.3.2 Arable arable grass (ley)
4.3.3 Arable setaside (bare)
4.3.4 Arable setaside (undifferentiated)
5.1.1 Grass improved intensive
5.1.2 Grass improved (hay/ silage cut)
5.1.3 Grass improved grazing marsh
5.2.1 Grass setaside
6.1.1 Grass rough (unmanaged)
6.1.2 Grass (neutral / unimproved)
7.1.1 Grass calcareous (managed)
7.1.2 Grass calcareous (rough)
8.1.1 Grass acid
8.1.2 Grass acid (rough)
8.1.3 Grass acid with Juncus
8.1.4 Grass acid Nardus/Festuca/Molinia
9.1.1 Bracken
10.1.1 Dwarf shrub heath dense (ericaceous)
10.1.2 Dwarf shrub heath gorse
10.2.1 Dwarf shrub heath open
11.1.1 Fen, marsh, swamp
11.1.2 Fen, marsh
11.1.3 Fen willow
12.1.1 Bog (shrub)
12.1.2 Bog (grass/shrub)
12.1.3 Bog (grass/herb)
12.1.4 Bog (undifferentiated)
13.1.1 Water (inland)
15.1.1 Montane
16.1.2 Inland bare despoiled
16.1.1 Inland bare semi-natural
17.1.1 Suburban/rural developed
17.2.1 Urban residential/commercial
17.2.2 Urban industrial
18.1.1 Supra-littoral rock
18.1.2 Supra-littoral shingle
18.1.1 Supra-littoral shingle (vegetated)
19.1.3 Supra-littoral dune
19.1.4 Supra-littoral dune shrubs
20.1.1 Littoral rock
20.1.2 Littoral rock with algae
21.1.1 Littoral mud
21.1.2 Littoral sand
21.1.3 Littoral sand with algae
21.2.1 Supra-littoral saltmarsh
21.2.2 Supra-littoral saltmarsh (grazed)
22.1.1 Sea
Appendix 2

Colour scheme for LCM2000 land cover subclasses

Source: Full LCM2000 dataset description published by the Centre for Ecology and Hydrology
<table>
<thead>
<tr>
<th>LCM2000 Subclasses</th>
<th>Class Number, LEVEL 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sea / Estuary</td>
<td>22.1</td>
</tr>
<tr>
<td>Water (inland)</td>
<td>13.1</td>
</tr>
<tr>
<td>Littoral rock</td>
<td>20.1</td>
</tr>
<tr>
<td>Littoral sediment</td>
<td>21.1</td>
</tr>
<tr>
<td>Saltmarsh</td>
<td>21.2</td>
</tr>
<tr>
<td>Supra-littoral rock</td>
<td>18.1</td>
</tr>
<tr>
<td>Supra-littoral sediment</td>
<td>19.1</td>
</tr>
<tr>
<td>Bog</td>
<td>12.1</td>
</tr>
<tr>
<td>Dwarf shrub heath</td>
<td>10.1</td>
</tr>
<tr>
<td>Open dwarf shrub heath</td>
<td>10.2</td>
</tr>
<tr>
<td>Montane habitats</td>
<td>15.1</td>
</tr>
<tr>
<td>Broad-leaved woodland</td>
<td>1.1</td>
</tr>
<tr>
<td>Coniferous woodland</td>
<td>2.1</td>
</tr>
<tr>
<td>Arable cereals</td>
<td>4.1</td>
</tr>
<tr>
<td>Arable horticulture</td>
<td>4.2</td>
</tr>
<tr>
<td>Non-rotational horticulture</td>
<td>4.3</td>
</tr>
<tr>
<td>Improved grassland</td>
<td>5.1</td>
</tr>
<tr>
<td>Setaside grass</td>
<td>5.2</td>
</tr>
<tr>
<td>Neutral grass</td>
<td>6.1</td>
</tr>
<tr>
<td>Calcareous grass</td>
<td>7.1</td>
</tr>
<tr>
<td>Acid grass</td>
<td>8.1</td>
</tr>
<tr>
<td>Bracken</td>
<td>9.1</td>
</tr>
<tr>
<td>Fen, marsh, swamp</td>
<td>11.1</td>
</tr>
<tr>
<td>Suburban/rural developed</td>
<td>17.1</td>
</tr>
<tr>
<td>Continuous Urban</td>
<td>17.2</td>
</tr>
<tr>
<td>Inland Bare Ground</td>
<td>16.1</td>
</tr>
</tbody>
</table>
Appendix 3

Simulation of uncertainty based on LCM2000 metadata

- Alternative realizations of the Derwent and Colsterworth squares and animated GIF of the Colsterworth square

(see attached CD for animation using Microsoft Powerpoint)
Alternative realisations of the LCM2000 based on per-pixel list data

All of the following maps use the colour scheme given in Appendix 2.

Twenty realisations of the Derwent square:
Twenty realisations of the Colsterworth square:
Evaluating uncertainty in classification within the Land Cover Map 2000 - Appendices

[Images of land cover maps]

245
Appendix 4

Glossary of abbreviations
Appendix 4: Glossary of abbreviations

ANZLIC: The Australia and New Zealand Land Information Council


CEH: The Centre for Ecology and Hydrology

FGDC: Federal Geographic Data Committee (US).

GI: Geographic Information

GIS: Geographic Information System

ISO: International Standards Organisation

LCM2000: The Land Cover Map 2000 produced by the Centre for Ecology and Hydrology

LCMGB: The Land Cover Map of Great Britain produced by the Centre for Ecology and Hydrology in 1990

MLC: Maximum likelihood classifier. A parametric multivariate statistical technique commonly used to classify multi-band spectral imagery.

Moran's I: A measure of spatial autocorrelation.
NCDCDS: The National Committee for Digital Cartographic Data Standards (US).

NIWT: National Inventory of Woodland and Trees produced by the Forestry Commission from aerial photography

PixA - PixE: Code used as the name for component parts of the Per-Pixel List (PPL). PixA being the land cover type identified by the per-pixel classification as occupying the most pixels within the parcel. PixB is the land cover type occupying the second largest number of pixels, and so on.

Pix1 - Pix5: Code used as the name for component parts of the Per-Pixel List (PPL). Pix1 being the percentage value of the parcel occupied by the PixA land cover type, and so on.

PolA - PolE: PolA is the land cover type with the highest spectral probability from the per-parcel classification. PixB is the land cover type with the second highest spectral probability from the per-parcel classification, and so on. This data is not included as standard metadata within LCM2000.

Pol1 - Pol5: Pol1 is the value for the spectral probability assigned to PolA by the per-parcel classification. Pol2 is the value for the spectral probability assigned to PolB, and so on.

PPL: Per Pixel List — a field of the metadata provided within the Land Cover Map 2000 giving the top five land cover types by area in each parcel as identified by individual pixel classification (see Table 2.2).
Appendix 5

Evaluating Object-Based Data Quality Attributes in the Land Cover Map 2000 of the United Kingdom

Paul Robinson, Peter Fisher, and Geoff Smith

Abstract
Standards that have been created for the reporting of data quality in spatial databases focus primarily on database level metadata, which does not address the issue of varying quality within a dataset. This issue may be advanced by the inclusion of metadata at an object level. The Land Cover Map 2000 (LCM2000) is a national database for the UK and contains a large amount of object-based quality metadata, which is reviewed. An analysis of uncertainty in the extent of three land cover types is carried out using a cumulative evidence method, utilizing a number of existing datasets, each purporting to represent the extent of the phenomenon in question. The output of this analysis is compared to the metadata in the LCM2000, in order to assess the usefulness of the metadata in understanding attribute accuracy. Results of this comparison are presented, showing that some of the object-based metadata gives a useful indication as to the certainty of classification.

Introduction
Spatial data will always contain inherent uncertainties and errors (Burrough and McDonell, 1998). There has been a great deal of work carried out in the last decade or more to decide how best to report the quality of digital spatial databases (NCDCDS, 1988; Guptill and Morrison, 1995; Foody, 2002; Joos, 2003). If such reporting is carried out in a constructive way then the user of the data will be empowered to decide if that data is suitable for their needs. Although, in practice, users may well not be interested in the uncertainties involved (Goodchild, 1998), there is growing pressure on data producers to provide information which is compliant with International Standardization Organisation (ISO) specifications on the quality of their products. Such information about the composition and construction of the data is called metadata. The widespread use of technologies such as GIS has increased the need for quality metadata, as digital spatial databases are not necessarily a final product themselves, as was the case with a paper map, but an input into analyses carried out by the user (Morrison, 1995; Smith and Fuller, 2002).

Various proposals for standardised metadata have been created by different organizations, with such aims as allowing easier comparison of datasets from different providers, which will form part of the decision a purchaser has to make between competing datasets. There is an increasing need for quality metadata at all levels from the dataset (or mapset) level, and increasing compatibility of datasets (Green and Bossonmaier, 2002). Such standards often, if not always, include information regarding data quality. The National Committee on Digital Cartographic Data Standards (NCDCDS) in the United States published an early example of such a standard created for digital spatial data (NCDCDS, 1988). The committee proposed five components for reporting spatial data quality, being lineage, positional accuracy, attribute accuracy, completeness, and logical consistency (see Morrison (1995) for detailed descriptions of these components). Later work by the Commission on Spatial Data Quality of the International Cartographic Association proposed two further components: semantic accuracy and temporal information (Guptill and Morrison, 1995).

Metadata standards created since the NCDCDS report have often only required positional accuracy information (Goodchild, 1995), while others have used various combinations of the data quality components outlined above. The Australia and New Zealand Land Information Council (ANZLIC) published their standard in 1998 adopting the same components as proposed by NCDCDS (ANZLIC, 2001, Green and Bossonmaier, 2002), as did the Federal Geographic Data Committee (FGDC, 1998). Technical Committee 211 of the ISO (ISO/ TC 211) created ISO19113 (Quality Principles), which additionally contains a section relating to temporal accuracy (Joos, 2003).

As already discussed, compliance with such standards serves the purpose of allowing easy comparison and increasing compatibility between datasets. The standards, however, are principally intended for use at the dataset level, and metadata at this level does not address the problem of quality varying within a large dataset, as is often the case (Smith and Fuller, 2002). This issue may be advanced by the inclusion of metadata at levels other than that of the dataset. Fisher (1993) discusses the potential benefits of including quality metadata at all levels from the dataset (or mapset) to the level of the individual object, suggesting examples of metadata that would be relevant at each level. Aalders (2002) envisages metadata reporting of quality information relating to homogeneous subsets of the dataset and Gan and Shi (2002) created a system to manage and update object level metadata that complies with FGDC standards for a topographic map series. The ANZLIC standard discusses the
potential for object level metadata, though it is not implicitly built into the reporting structure. Data quality metadata is included at the object level within the 1:500000 topographic map of Australia (Geoscience Australia, 2003). The ISO19115 (metadata) standard does give the opportunity for quality reporting at a level specified by the user. This could be at the object level, a regional level, or for the whole dataset, though any level below that of the dataset is optional (ISO, 2000), and the level of reporting has to be specified if it is more detailed than the dataset level (Joos, personal communication, 2003). In the case of land cover mapping, which is the focus of this paper, it could be possible and relevant to include all of the previously mentioned components of spatial data quality as object-based metadata. That is with the possible exception of semantic accuracy, which deals with the suitability of the classification model employed within the dataset.

Goodchild (1995, p. 76), when discussing an example of a classified soil map, suggested, "we might go so far as . . . to tag each polygon with an estimate of the proportions of various classes present within it"; in other words, use object-based metadata to record information with information on data quality. This is a step towards recording a probability for all possible classes at all pixels in a land cover map, rather than a single, inflexible land cover class (Foody, 1996; Mather, 1999). By using this approach the user is supplied with adequate metadata to decide whether or not the data is appropriate for their needs in specific terms, such as for the specific study area in which the user is interested, rather than just in general terms as with database-level quality statements. Database-level metadata regarding quality is still potentially useful in some circumstances such as for semantic accuracy, although this could be reported at the category level. Database level quality statements have been successfully used to model issues of accuracy in DEMs (Fisher, 1991; Hunter and Goodchild, 1997). Even in DEM work, Fisher (1998) has recently called for distributed models of error and meaningful object level metadata for such a distributed model of quality, which should lead to more spatially responsive models.

In this paper we evaluate object-based data quality measures reported with the Land Cover Map 2000 (LCM2000) of the United Kingdom (UK) with respect to assessing attribute uncertainty. In the next section, the metadata reporting in that dataset is reviewed. Ancillary datasets used in the evaluation are then introduced, and the methodology used in analysis reported. Results are then presented, and concluding comments are made.

The Land Cover Map 2000

LCM2000 is a national database covering the whole of the UK, published by the Centre for Ecology and Hydrology (CEH). It was produced by classification of satellite image data, with an image segmentation process producing classified vector parcels as opposed to the more usual raster pixels (Smith and Fuller, 2002). This segmentation was carried out by an algorithm and identified spectrally uniform areas within the image. The algorithm used a region growing procedure from seed points having already identified edge features to ensure that seed points were not located on boundaries (Fuller et al, 2002). Within the database, users may acquire a considerable amount of object-based metadata, including data quality information, giving detail as to the production history of each parcel (see Table 1 for details). To the author's knowledge, the LCM2000 is the first national spatial database to contain such a level of object-based metadata especially relating to attribute uncertainty.

In the production of LCM2000, two classifications were carried out. One classified each individual pixel, and so is referred to here as the per-pixel classification. The other was carried out on each parcel of land once it had been created using the image segmentation algorithm. To classify the parcel, the averaged spectral response of the core pixels was used as input to a maximum likelihood classifier. Pixels around the edge were not included in the average of the parcel where the parcel was large enough to retain a specified minimum number of pixels after the removal of those on the edge. Using only pure core pixels in this classification reduced the impact of edge effects on the classification (Smith and Fuller, 2002). From this parcel classification, the spectral probability for each class is then reported in the metadata, being an indication of how close the average spectral response of the parcel is to that of training data. Results of the pixel-based classification within the parcel (see the PerPixList in Table 1) are summarized as the percentage of the pixel that had been attributed to each land cover category. The five categories with the highest percentages were stored and are available within the metadata. These are termed the per-pixel data and are completely independent from the classification of the parcel as it was carried out in an entirely separate process. The parcel classification was subjected to a knowledge-based corrections process (KBC), so that certain errors were removed and some fine separations of categories included (e.g., acid, neutral, and calcareous grass: Fuller et al, 2002).

As an initial stage in assessing the potential usefulness of the object-based metadata available within LCM2000, with respect to understanding attribute accuracy within the map, certain categories within the database were compared with what we have termed a cumulative evidence analysis. This analysis took evidence from other databases that the category is present and calculated a sum of the number of those databases that agree on the presence of the land cover type for all locations within the study area. We compared the extent of woodland, urban areas, and water in different mappings of those phenomena, and computed the number of times a location was identified as the named cover type with the support for the proposition in the metadata. The study area used in the work reported here was within the

Table 1. Object-Based Metadata Available with LCM2000 (Adapted from CEH, 2002)

<table>
<thead>
<tr>
<th>Title</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SegID</td>
<td>Unique identifier for each parcel.</td>
</tr>
<tr>
<td>TotPixels</td>
<td>Total number of pixels in each parcel.</td>
</tr>
<tr>
<td>CorePixels</td>
<td>Number of pixels in core area of the parcel used in per parcel classification.</td>
</tr>
<tr>
<td>BHSubVar</td>
<td>Hierarchical code detailing the dominant land cover type for each parcel, giving broad habitat and subclass.</td>
</tr>
<tr>
<td>BHSub</td>
<td>Hierarchical code detailing the dominant land cover type for each parcel, giving broad habitat and subclass. From same production process as BHSubVar.</td>
</tr>
<tr>
<td>OpHistory</td>
<td>Processing history of each parcel including input images, spectral probability from maximum likelihood classification of parcel and the number of KBC rules applied. Includes quality data on lineage, completeness and corrections made relating to logical consistency.</td>
</tr>
<tr>
<td>PerPixList</td>
<td>Top five land cover types by area as identified by individual pixel classification, giving percentage of parcel allocated to each type. Produced by maximum likelihood classification of each individual pixel in the image. Not related to BHSub or BHSubVar. Includes quality data on attribute accuracy.</td>
</tr>
</tbody>
</table>
boundary of the National Forest in the English midlands (Figure 1), which covers an area of more than 500 km². The National Forest was established in 1995 by the UK government, with the specific aim of dramatically increasing the amount of woodland cover, and so bringing about a range of environmental and social benefits.

**Input Databases for Cumulative Evidence Analysis**

Table 2 outlines the datasets used in the analysis of attribute uncertainty, and clearly shows the variety of different intended scales and purposes in their creation. As they are to be used as evidence of the presence of specific land cover types, it would be preferable to use datasets that have a shared ontology, but it is unlikely that many mappings of the same area at the same scale for the same purpose at the same time (a fully shared ontology) have ever occurred in more than a very limited number of controlled experimental situations (Edwards and Lowell, 1996; Middelkoop, 1990). It certainly has not occurred for the land cover of Great Britain.

### Table 2. The Sources and Characteristics of Datasets Used as Inputs to the Cumulative Evidence Analysis

<table>
<thead>
<tr>
<th>Dataset Title</th>
<th>Copyright Owner/Originating Organisation</th>
<th>Resolution/Associated Scale</th>
<th>Availability</th>
<th>Data Format</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bartholomew’s dataset</td>
<td>Collins Bartholomew</td>
<td>1:200000</td>
<td>Licence agreement</td>
<td>Vector</td>
<td>Enhanced and updated version of 1:250000 OS data.</td>
</tr>
<tr>
<td>Meridian dataset</td>
<td>Ordnance Survey</td>
<td>1:250000</td>
<td>Licence agreement</td>
<td>Vector</td>
<td>Originated from OS Strategy, a 1:250000 product.</td>
</tr>
<tr>
<td>Land Cover Map of Great Britain (1990; LCMGB)</td>
<td>Centre for Ecology and Hydrology</td>
<td>1250 m² Minimum mapping unit (MMU)</td>
<td>Licence agreement</td>
<td>Raster</td>
<td>25m raster grid created from satellite imagery using maximum likelihood classifier</td>
</tr>
<tr>
<td>Land Cover Map 2000 (LCM2000)</td>
<td>Centre for Ecology and Hydrology</td>
<td>5000 m² (MMU)</td>
<td>Licence agreement</td>
<td>Vector</td>
<td>Parcel-based dataset created from segmented satellite imagery</td>
</tr>
<tr>
<td>National Inventory of Woodland and Trees (NIWT)</td>
<td>Forestry Commission</td>
<td>1:25000–1:50000 maps: 2 ha MMU</td>
<td>Licence agreement (<a href="http://www.forestry.gov.uk">www.forestry.gov.uk</a>)</td>
<td>Vector</td>
<td>1:25000 aerial photographs and ground survey</td>
</tr>
<tr>
<td>Ancient Woodland Inventory (AWI)</td>
<td>English Nature</td>
<td>1:25000 maps: 2 ha MMU</td>
<td>Public Domain (<a href="http://www.englishnature.org.uk">www.englishnature.org.uk</a>)</td>
<td>Vector</td>
<td>Various paper maps, with base maps from the 1920s</td>
</tr>
<tr>
<td>Census Urban Wards</td>
<td>EDINA</td>
<td>1:2500–1:50000 maps</td>
<td>Licence agreement</td>
<td>Vector</td>
<td>Based on census wards defined as urban or rural</td>
</tr>
</tbody>
</table>

The ideal situation of a fully shared ontology between the inputs is therefore not possible, but the aim of the work reported here was to utilize as many extant datasets as possible that report aspects of land cover over a large area. The premise being that the more datasets that are in agreement about the presence of a particular phenomenon, the lower will be the uncertainty of that phenomenon occurring at that location. We refer to this as a cumulative evidence analysis. As noted above, the datasets available are heterogeneous in their ontology, which will obviously lead to differences, but all purport to represent the extent of the phenomenon in question within the study area. In these aspects the ontology is shared. Most were also created or last updated over a similar timescale, with the National Inventory of Woodland and Trees (NIWT) and the Ancient Woodland Inventory (AWI) both completed in 2000, OS Meridian Data (GeoBusiness Solutions, 2004) being updated in 2001 and Bartholomew’s Dataset (Collins Bartholomew, 2003) undergoing a major enhancement in 1998. The Urban Settlement project was started in 2001 and LCM2000 is created from satellite imagery dating from the late 1990s. The final two datasets are slightly older with the census wards being those of the 1991 census and the Land Cover Map of Great Britain (LCMGB) created from satellite imagery from 1988 to 1990. There is therefore an issue with currency of data, which cannot be avoided, leading to the possibility that land cover can change between production of the different datasets. A majority of such change is likely to be over small areas and so would be detected by those datasets with a fine resolution, but is less likely to affect the coarse resolution products. Such a change would lead to some of the datasets designating an area as the target land cover class and some would not, so there would be some evidence for that phenomenon at that location, but it would not be certain as not all datasets are in agreement. This is still an indication of uncertainty from the analysis, which is the aim, given the lack of a definitive, totally accurate map.

Resolution is another property of the datasets that is not shared (Table 2), nor is the purpose for creating the datasets, with some products aiming to create an exhaustive map of land cover (LCMGB, LCM2000), some being general topographic...
datasets (Meridian, Bartholomew’s), one being based on administrative boundaries (census wards), and others being very specific analyses of the extent of particular phenomena (AWI, NIWT, Urban Settlements).

As far as possible it has been ensured that the datasets are not directly derived from one another. Bartholomew’s Data was originally sourced from Ordnance Survey OS 1:250000 mapping, but has since been updated from photographic and survey evidence by the publishers (GILL, personal communication, 2003). AWI is also sourced from OS mapping, but at 1:25000 scale and using 1920s maps as the base for its production. The Urban Settlement boundaries were sourced from current OS 1:10000 maps with other datasets being produced independently from aerial photography and satellite imagery.

Method
All datasets for each category were converted into binary presence-absence maps on a 5 m grid and added together to calculate the number of times the phenomenon in question was present. The output was a map that gave a value to every location in the study area between zero (all datasets agreed that the category is absent) and the total number of datasets (all agreed that the category is present), which is four in the case of open water and six for woodland and urban areas. This map of values was taken to be a representation of the level of certainty in the presence of that category. Middelkoop (1990) and Edwards and Lowell (1996) utilized a similar method, of combining different realizations or interpretations of the same set of objects, thus building a map of the level of agreement within those realizations. In both cases they were analyzing boundary uncertainty using multiple manual interpretations of images. Molenaar (1998) also used such a method and was able to calculate a fuzzy set membership value for all areas, thus indicating the reliability of the classification. Here, we simply took the ranked value from the overlay operation as a relative measure of certainty of the mapped cover type (Table 3).

These maps were then compared with both the per-parcel and the per-pixel classification metadata from LCM2000, in order to see if patterns of uncertainty were similar between the two analyses. A list was generated of all the LCM2000 parcels that overlapped with areas of each cumulative evidence value. In the comparison any parcels that had been subjected to KBCs were removed from the per-parcel analysis, as the corrections would render relevance of the spectral probability value questionable. The parcels subject to KBCs were not removed from the per-pixel analysis as this was derived from a separate classification procedure and so was not affected by the KBCs. Parcels that had been filled manually due to such problems as cloud cover in the satellite images were removed however, as these contained no per-pixel data. The number (n) of parcels overlapping with each value was counted and the per-pixel and per-parcel metadata within those parcels was analyzed.

Results
Metadata from Per-Pixel Classification-Woodland
The outcome of the assessment of uncertainty regarding the presence of woodland (Figure 2) showed that total agreement in the presence of woodland covered just 0.83 percent of the study area (Table 3). This is considerably less than the lowest area given by any of the individual data sets, which was 2.61 percent of the study area (Table 3). The area allocated value 4 and above, meaning that more than half of the data sets agreed that woodland was present, still only accounts for 2.57 percent, illustrating a considerable level of disagreement between the data sets.

The per-pixel data was reviewed to give a percentage of each parcel that had been designated as a woodland category. This percentage was compared with the values generated by the cumulative evidence analysis. Parcels with each value (0–6) were selected, and the attribute data analysed to assess the percentage of the parcel designated as woodland by the per-pixel classification. This analysis therefore reflects parcels

![Figure 2. Map of the National Forest showing total agreement on the presence of woodland or value 6 (black), total agreement on the absence of woodland, or value 0 (white) and non-agreement, or values 1–5 (grey).](image)

### Table 3. Proportion of Study Area Classified as Each Land Cover Type by the Datasets Used as Inputs to the Cumulative Evidence Analysis

<table>
<thead>
<tr>
<th>Dataset Title</th>
<th>Wood %</th>
<th>Urban %</th>
<th>Water %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bartholomew’s dataset</td>
<td>3.01</td>
<td>7.76</td>
<td>0.80</td>
</tr>
<tr>
<td>Meridian dataset</td>
<td>2.61</td>
<td>13.13</td>
<td>0.77</td>
</tr>
<tr>
<td>Land Cover Map of Great Britain (1990; LCMGB)</td>
<td>4.92</td>
<td>19.16</td>
<td>0.46</td>
</tr>
<tr>
<td>Land Cover Map 2000 (LCM2000)</td>
<td>8.70</td>
<td>15.82</td>
<td>1.04</td>
</tr>
<tr>
<td>National Inventory of Woodland and Trees (NIWT)</td>
<td>9.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ancient Woodland Inventory (AWI)</td>
<td>2.86</td>
<td>9.12</td>
<td>17.59</td>
</tr>
</tbody>
</table>

### Table 4. Output of Cumulative Evidence Analysis for Woodland, Showing the Percentage of the National Forest Area Allocated to Each Value

<table>
<thead>
<tr>
<th>Ranked Value</th>
<th>Proportion of National Forest in Analysis</th>
<th>Proportion of National Forest in Urban Areas Analysis</th>
<th>Proportion of National Forest in Open Water Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0.83%</td>
<td>2.54%</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.79%</td>
<td>2.81%</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.95%</td>
<td>2.08%</td>
<td>0.17%</td>
</tr>
<tr>
<td>3</td>
<td>1.84%</td>
<td>3.06%</td>
<td>0.23%</td>
</tr>
<tr>
<td>2</td>
<td>2.54%</td>
<td>6.72%</td>
<td>0.29%</td>
</tr>
<tr>
<td>1</td>
<td>8.64%</td>
<td>22.36%</td>
<td>1.11%</td>
</tr>
<tr>
<td>0</td>
<td>84.44%</td>
<td>60.44%</td>
<td>98.20%</td>
</tr>
</tbody>
</table>
created by the intersection process, and so the number of parcels is not the same as in the LCM2000, and the same parcel (and so the same metadata) can appear in more than one value. Data for each of the values is summarized in Figure 3.

None of the per-pixel datasets were normally distributed (verified using an Anderson-Darling test), so a nonparametric one-sample sign test was used to calculate the 95 percent confidence intervals (CIs), based on median values. This shows that values 0 and 1 are indistinguishable and that there is a slight overlap in confidence intervals between values 3 and 4, and values 4 and 5. Confidence intervals exhibit no overlap at the 99 percent level, other than between the two lowest values, therefore, per-pixel values could predict the ranked value of a location with this level of confidence.

Metadata from Per-Pixel Classification-Urban Areas

Urban areas were analysed using the same process as described above. The proportion of the study area designated the different values is given in Table 4, and the map of the output is shown in Figure 4. Total agreement in the presence of urban areas (2.54 percent; value 6, Table 4) is again considerably less than the lowest area given by any individual dataset, which is 7.76 percent (Table 3). The area allocated 4 or above, accounts for 7.43 percent of the study area.

Figure 5 shows that there is again a clear relationship between the cumulative evidence values and the per-pixel percentages. Despite the range of percentages being zero to 100 for each value, the 95 percent CIs display no overlap other than between values zero and one (Figure 5). This is also true for CIs at the 99 percent level.

Metadata from Per-Pixel Classification-Open Water

The results from the analysis of open water (Table 4, and Figures 6 and 7) are much less satisfying than those from woodland and urban areas. This is because there is not very much open water within the study area (Table 4) leading to very small sample sizes for locations allocated values of three or four in the cumulative evidence analysis. For this reason the results cannot be seen as in any way conclusive, but they do still exhibit a pattern of increasing per-pixel
percentages in relation to higher values from increasing cumulative evidence.

Only 0.17 percent of the study area has total agreement in the presence of open water (Table 4). The proportion of the study area allocated a value of three or four is 0.40 percent, which is slightly less than the smallest proportion designated as water by any of the individual datasets.

Again the two lowest values are indistinguishable from one another and there is some overlap between the 95 percent confidence intervals (Figure 7). There is no overlap of confidence intervals at the 85 percent level other than the two lowest values, which still have a range of zero at this level.

**Discussion**

**Metadata from Per-Pixel Classification**

Figures 3, 5, and 7 show that, for each of the three land cover types reviewed, the metadata from the per-pixel classification gives a useful indication as to the uncertainty regarding the classification at a land parcel level. As certainty of the evidence for each land cover type increases, the percentage of the parcel designated as that category by the per-pixel classification also increases. Demonstrating that the metadata from the per-pixel classification reflects confidence from independent datasets that the category is present.

While Figures 3, 5, and 7 all display a similar trend in the relationship between the uncertainty analysis and the metadata from the per-pixel classification, the detail within them is quite different. In the cases of woodland and urban (Figures 3 and 5), the increase in the mean percentage is close to being linear, whereas the increase in the graph of water is not (Figure 7). This will be affected by the smaller sample size at higher values for water, and it is possible that the relationship looks most linear in the case of urban areas, because there are a greater number of samples at higher values.

For all of the three land cover types reviewed here, the per-pixel percentages of parcels intersecting areas allocated value 0 and 1 are statistically indistinguishable as the full range of the 95 percent confidence intervals is zero. This is still the case even at the 85 percent confidence level, at which point there is no overlap in the confidence intervals between any of the other values. The implication of this is that when dealing with datasets that have such heterogeneous resolutions and potential mis-registrations, a single
piece of evidence as to the presence of a phenomenon has to be deemed unreliable. Interestingly, in the case of woodland the smallest dataset in terms of spatial coverage, represents 2.61 percent of the study area (Table 1). At the point where more than half the datasets agree (those given a value of four or more) 2.57 percent of the study area is represented (Table 4), so the total area is very similar. This is also the case with urban areas where the smallest dataset covers 7.76 percent of the area and those regions of value 4 and above cover 7.43 percent. When only using four datasets, as with water, the area covered is not quite so similar. It is unclear if this is explained by the smaller number of datasets or the smaller areas covered, leaving potentially greater impacts from mis-registration.

The graph of urban areas (Figure 5) is the only one to have a range for percentages of 0 to 100 for all values. The other graphs show a range from 0 to 4 for most of the values. This would lead to higher sample sizes within the higher values and probably to lower average percentages for each value, though it is felt unlikely to greatly change the general pattern of the results (Figures 3, 5, and 7).

**Metadata from Per-Parcel Classification**

It has been suggested that the spectral probability from a maximum likelihood classifier can be used as a measure of classification uncertainty (Goodchild et al., 1992; Fisher, 1994; van Deusen, 1995; van der Wel et al., 1998; de Bruin and Gorte, 2000; Foody, 2002). Normally, only the most likely class is reported, which means that not all of the information created within the classification process is recorded (Foody et al., 1992). It follows then that reporting the probability, even just that of the most likely class, as is the case with the per-parcel metadata in LCM2000, can give a measure of the spatial distribution of uncertainty (Foody, 2002). This is moving towards a soft classification approach and is how the per-parcel metadata is reported in the LCM2000, with the probability of the most likely class included.

If the per-parcel metadata from LCM2000 gives an indication of uncertainty in the classification of the parcel, then there should be a predictable relationship between it and the cumulative evidence values. Two relationships could be predicted. First, in those parcels that have been classified as the target land cover type the spectral probability of those parcels should increase as the certainty value increases. Second, the opposite relationship could be predicted in parcels that were not classified as the target cover type. Figures 8, 9 and 10 show that no such pattern is evident. Any trends that may be present are too small to be significant. This suggests that the spectral similarity between a parcel and the training data does not indicate uncertainty in the classification (Smith and Fuller, 2002). The production process of LCM2000 could be, at least in part, responsible for this, as the spectral probability figure relates to the average spectral response of the parcel and not of an individual pixel, and so is partly a function of the segmentation algorithm used and production of the parcels. Also, the reported spectral probability for the most likely land cover type relates to a land cover subclass. If these probabilities were amalgamated to a higher level of categorisation, then the relationship with the cumulative evidence values is likely to be very different.

The pattern of responses in Figures 8, 9, and 10 would imply that a set of input datasets with a fully shared ontology would have little impact on the relationships displayed in these graphs. Percentages display very similar ranges across all the values and therefore having larger areas allocated to higher values is unlikely to greatly alter the ranges of percentages within each value.

**Conclusion**

The LCM2000 is a pioneer of object level reporting of data quality metadata. While standards have specified information for database level metadata, little has been written on object-based metadata. Therefore, the measures used in the LCM2000 have been developed in a rather ad hoc fashions. This analysis shows that some of the object-based metadata (per-pixel lists) relating to attribute accuracy available with the LCM2000, at least for certain categories, gives a useful indication as to the uncertainty of classification at a land parcel level. Other measures may not (spectral probability), meaning that users can take the spatial variation of uncertainty in the data into account. While it may be intuitively appropriate to exploit the spectral probability from the maximum likelihood classifier as an indicator of attribute accuracy, the statement here suggests that this may not be the case (Smith and Fuller, 2002). Associating the results of an entirely independent classification procedure to the map at a land parcel level, gives users of the LCM2000 a tool by which they can develop an understanding of the impact of categorical uncertainty on their work. It was not the aim of this analysis to make recommendations as to how this metadata may be utilized, that is an issue that will be addressed in future work.

**Acknowledgments**

The authors would like to thank the University of Leicester for financial support and the Centre for Ecology and Hydrology and the Forestry Commission for provision of data, as well as two reviewers for their constructive comments. Further thanks are extended to the National Forest Company and particularly Annette McGrath for the provision of data and assistance.

**References**


Mather, P.M., 1999. Land cover classification revisited, Advances in remote sensing and GIS analysis (P.M. Atkinson, and N.J. Tate, editors), John Wiley and Sons, Chichester.


