Methods for Estimating the Accuracy of Per-Pixel, Per-Parcel and Expert Visual Classification of High Resolution Optical Satellite Imagery

Neil Stuart †, Tom Jaas, Ioannis Zisopoulos and Karin Viergever
Institute of Geography, University of Edinburgh, Drummond Street Edinburgh EH8 9XP.

Abstract. We describe methods for collecting appropriate quantities and types of reference data for validating classifications of high resolution satellite data. We use the example of collecting reference data to test classifications of 1m spatial resolution IKONOS data for an open woodland savanna in Central America. Reference data was collected in the field by GPS survey to ensure the purity and representativeness of the ground areas and a precise matching between the ground data and the corresponding image pixels. The image is then classified by three methods: by automatic per-pixel maximum-likelihood (ML), by automatic per-parcel nearest neighbour and by a visual classification by experienced image interpreters. We find that the per-parcel classifier achieves higher accuracy than the per-pixel ML classifier for all the required land cover classes. The overall accuracy for the per-parcel classifier is 82% (producer accuracy range: 47-95%, k=0.73) compared to 57% (range: 36-70%, k=0.5) for ML. The classification by expert visual interpretation yields an overall accuracy of 96% (range: 89-100%, k=0.95). The per-parcel classification exceeded the minimum accuracy requirement of 70% for two of five required land cover classes and approached the target of 85% suggested for the overall accuracy required in natural resource mapping. We conclude that a per-parcel classifier can achieve an acceptable standard of accuracy for some of the savanna land cover classes, but that further work is needed to improve the classification of smaller groups trees and sparse woodlands.

Since visual classification is still commonly used in developing countries for classifying imagery and in some cases is the desired output that an automated classifier seeks to reproduce, we developed a means to measure the stability or reliability of a visual classification. We estimate the average accuracy of a series of manual, visual classifications of the same image by different interpreters, by comparing each to the agreed ‘master’ classification by an expert interpreter. The result shows which map features are frequently classified correctly (or not) by different interpreters and according to their level of expertise. This information allows further training to focus on these classes.

Keywords: accuracy comparison, per-parcel, visual interpretation, reference data, IKONOS.

1. Introduction

Despite higher resolution (HR) remotely sensed imagery becoming more affordable for natural resources management, this increase in spatial resolution has not always led to improved accuracy of land cover classification. There are several reasons why the accuracy:cost ratio may not be improved by using HR imagery. If land cover types are not separable by their spectra, they may remain inseparable in optical imagery of higher spatial resolution. When ground features are represented by more pixels, the within-feature variance may increase and make it difficult to discriminate the feature from its surroundings [1].

As well as the higher cost of HR data, more fieldwork may be involved to collect the required quantities of ground control and reference data to verify the image classification. The use of HR imagery increases both the volume and positional accuracy requirements for the ground reference data for training and testing the classification. Greater positional accuracy is needed to ensure precise matching between the image pixels and the ground reference conditions. As the resolution of the imagery increases, proportionately fewer pixels in an image will have ground reference data collected for them, unless a greater amount of resource is...
committed to the ground survey. As the costs of ground surveys increase, the trend is in fact to reduce the amount of ground effort required and seek other sources of control and reference data [2]. The smaller ground extent that is imaged in one scene of HR optical data also increases the likelihood of a substantial proportion of an image being obscured by cloud cover, or not containing enough clearly visible ground control features to enable a high quality georectification, in both cases leading to a poorer quality of result. These problems have sometimes led to poor accuracy when conventional automatic image classifiers are used on HR data [3]. This has prompted researchers to consider different approaches for classifying HR data [4]. Two contrasting possibilities are to use automated per-parcel classifiers and manual visual interpretation.

Particularly in developing countries, the resources available to undertake the classification of remotely sensed imagery also need to be considered. These include the availability of data and software and also the human resource capacity such as the number of skilled operators of image processing software or numbers of staff with expertise in interpreting remotely sensed images. Since many types of HR data can be presented in ways that make them appear similar to aerial photography, there has been interest in using simpler manual methods of visual interpretation to classify such imagery. Visual interpretation has a relatively low resource input, since it does not require advanced image processing skills and can be undertaken entirely manually. Human interpreters can often cope better than automatic classification methods with problems such as haze, cloud cover and shadow. Nevertheless, visual interpretation remains a time-consuming task and finding sufficient trained interpreters to undertake interpretation of extensive amounts of imagery remains a major limitation of the method [5]. Whether it is used as the classification technique of choice, or as a source of reference data for checking more automated methods, visual classifications always include subjectivity due to the skills and experience of the interpreter, the features that are being identified and the clarity or vagueness of the boundaries between the features. Some means for measuring this variability between interpreters needs to be included in any visual interpretation [6].

An alternative and more technologically sophisticated approach is the use of semi-automatic ‘per-parcel’ classifiers. The ‘object-oriented’ nature of these classifiers offers the possibility, once the classifier is correctly trained, for automatically extracting features from imagery that would be visible to the human eye. If these classifiers could achieve an acceptable standard of classification accuracy, they could in principle be used to classify large extents of satellite data rapidly. However, this capacity remains to be proven for a variety of different global land cover types and sensor data [7].

1.1. Objectives
The objectives of this work were therefore to firstly evaluate the accuracy of the per-parcel classifier for creating land cover classifications, secondly to compare this with conventional per-pixel classifier and with a manual visual classification. Finally, we sought to devise a means to measure the variability (or stability) of visual classifications produced by interpreters with different levels of experience with this imagery. The choice between using more sophisticated automated classifiers and simpler manual methods has significant implications for the types of resource investments made in software and skills training. For this study we chose the context of classifying IKONOS imagery for the purposes of producing land cover mapping in a developing country, where both types of resources are scarce.

2. Study site and data
2.1. Rio Bravo Conservation Area in Belize, Central America;
Figure 1 shows the location of the study area in the Rio Bravo Conservation and Management Area (RBCMA) of NW Belize, Central America, near to the Hillbank Research Station (17° 36’N, 88° 41’W). This study area has been used for previous research classifying savanna land cover types using optical LANDSAT data [8] and for estimating above ground biomass of the woodland areas using airborne inSAR data [9]. The area is relatively characteristic of the broader land cover type, comprising a mosaic of grasslands, palms and woody species that occur in patches with pine and oak woodland [10]. Some of these vegetation types are illustrated in figure 2. Savannas are a globally important, covering 40% of the tropics and supporting approximately 20% of world population, predominantly in developing nations [11,12].
2.2. IKONOS data
An IKONOS image with less than 10% cloud cover, covering the savanna south of Hillbank and west of the East Gate, boundary of the RBCMA, was acquired on the 7th of March 2007. The 4-metre multispectral image was pan-sharpened with the 1-metre panchromatic data in order to obtain maximum advantage for visual feature interpretation. The acquired image had been provisionally geo-referenced by the data provider to the Universal Transverse Mercator (UTM) projection and WGS84 datum.

Fig. 1: Location of the field site within the Rio Bravo Conservation Management Area of NW Belize. Savanna extents are derived from [13].

Fig 2: Examples of the vegetation types to be classified within the study area: (a) Pine woodland; (b) Palmetto palm thicket and (c) (foreground) Grass-dominated open areas with scattered pines, palmetto and woody shrubs.

2.3. Ground survey data collection
Ground-truth surveys remain an essential component of any accuracy assessment for classified satellite imagery [14]. The ground survey was performed to obtain precise point, line and polygon data for (i) confirming geo-rectification of the satellite imagery, (ii) creating training areas and deriving spectral signatures for a maximum likelihood classification of the image and (iii) creating a dataset for assessing the accuracy of the ML, per-parcel and the visual interpretations of the IKONOS image. The required data was collected in the field with a Trimble GeoXHTM GPS receiver and differentially corrected to sub-metre accuracy using data collected simultaneously from a Trimble 4000TM SSI base station.

Reference data for the pine woodlands and open grassland areas were recorded as points collected randomly by traversing through these areas following a stratified sampling approach, while the more impenetrable palmetto palm and wetland reed classes were initially recorded as lines or polygons by
traversing the perimeter of the feature. A series of points were then randomly selected from within these polygons after the data had been post-processed, to form a set of reference data for each class. For testing the classification of pine woodlands, it was necessary to modify the reference data to ensure that only points collected directly under tree crowns were used. Some of the ground data points that had been recorded as areas of pine woodland, had in fact been collected while walking in the gaps between the tree crowns, where the GPS signal was stronger. Because of the high spatial resolution of the imagery, some of these gaps were imaged as pixels of the grassland cover type and so such points had to be omitted from the reference data set. This illustrates the need to be mindful of the resolution of the imagery when collecting reference data. Once all data sets had been prepared, they were separated into two sets for training and for testing the classifications.

2.4. ‘Master’ visual interpretation

In order to assess the variability in visual interpretations of the IKONOS imagery by a series of different interpreters, a detailed and rigorous interpretation was first carried out on a subset of the IKONOS image, covering approximately 2.1 x 1.5 km. The interpretation was undertaken by two experts, who both had a detailed first-hand knowledge of the ground area. After interpreting the image separately, the two interpretations were then discussed, differences were resolved and the two merged to create an agreed ‘master’ interpretation. This data set was then used for two purposes; firstly the accuracy of this expert interpretation was evaluated by testing it against the same reference data used to evaluate the automatic classifications. Secondly, it was then accepted as a reference data set and used to evaluate the variability of the interpretations made by a panel of image interpreters, as will be described in a later section.

3. Classification procedures

3.1. Per-pixel and Per-parcel classifiers.

A supervised maximum likelihood (ML) classification of the image was undertaken using ERDAS Imagine™ software. Training areas were obtained from the training areas subset of the ground survey data. Polygonal training areas were created by a standard method provided by the software for ‘seeding’ and growing regions of similar spectral properties. Signatures for a series of provisional land cover types were assessed for their separability, using the JM distance criteria. Some classes that had significant overlap were then merged. While open grassland and canopy shadow had the highest JM divergence index with the other classes (>1300), the mixed-vegetation and pines only achieved ‘fair’ inter-separability (1200-1299); reeds and open grassland were poorly separable (<1199). Classification was then performed into five classes (mixed vegetation, pine, open grassland, reeds and palms).

Per-parcel classification of the IKONOS image was undertaken using Definien’s Ecognition™ software. Segmentation of the image was performed using all four bands (red, green, blue and NIR). A series of experiments varying the segmentation parameters were performed until an acceptable result was obtained in which the features extracted matched reasonably well with the land cover features that would be extracted by eye. Settings for this segmentation were: scale parameter =30; colour/shape ratio = 0.9/0.1; smoothness/compactness ratio =0.9/0.1. A supervised nearest neighbor classification was then performed, using training data for five classes (dense forest, pine, open grassland, reeds and palms) selected from the common ground reference data.

3.2. Visual classification procedures

30 people agreed to undertake a visual interpretation of an extract of the IKONOS image. The interpretations were conducted in a supervised workshop at the University of Belize in Belmopan, in April 2007. Among these 30 people were frequent remote sensing users, senior scientists, GIS specialists, park rangers, students and other professionals and government employees. Each person completed a short questionnaire in which they reported their level of familiarity with high resolution optical imagery. The differing knowledge levels of the interpreters was a vital element for this study, as we wanted to analyse what types of prior knowledge were needed for visually classifying a 1m spatial resolution image. Interpreters were familiarised with the data by first showing them examples of the different land-cover types and allowing them to identify examples of these features on the IKONOS image.
A different subset of the IKONOS image, approximately 2.1 x 1.5 km was then given to each interpreter. One half of the group were asked to perform a classification using a true colour composition image [Red - Green - Blue bands], while the other 15 were asked to classify a pseudo colour image [Red - Green - Near Infrared bands]. Detailed guidelines were given to the participants, such as the number and types of land cover classes and the minimum mapping unit they should try to identify. Each of the 30 visual classifications was then scanned and the polygons that had been drawn by each interpreter were digitised in ArcMap. Each shapefile was then transformed to raster and the Raster Calculator of the Spatial Analysis extension used to analyse similarities and differences between each of the interpreted maps and the ‘master’ classification that had been created earlier and served as ‘truth’ for these comparisons.

4. Results

4.1. Accuracy assessments of the three classifications

Tables 1-3 present the accuracy assessments for the maximum likelihood (ML) classification, the per-parcel classification and the visual classification of the IKONOS imagery.

Table 1: accuracy assessment for the ML classification

<table>
<thead>
<tr>
<th>Land-cover class</th>
<th>Mixed vegetation</th>
<th>Pine</th>
<th>Open grassland</th>
<th>Reeds</th>
<th>Palmetto</th>
<th>Rows total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixed vegetation</td>
<td>131</td>
<td>147</td>
<td>8</td>
<td>7</td>
<td>104</td>
<td>397</td>
</tr>
<tr>
<td>Pine</td>
<td>12</td>
<td>322</td>
<td>233</td>
<td>65</td>
<td>97</td>
<td>729</td>
</tr>
<tr>
<td>Open grassland</td>
<td>0</td>
<td>16</td>
<td>722</td>
<td>14</td>
<td>7</td>
<td>759</td>
</tr>
<tr>
<td>Reeds</td>
<td>0</td>
<td>7</td>
<td>30</td>
<td>293</td>
<td>3</td>
<td>333</td>
</tr>
<tr>
<td>Palmetto</td>
<td>103</td>
<td>404</td>
<td>173</td>
<td>33</td>
<td>498</td>
<td>1211</td>
</tr>
<tr>
<td>Columns total</td>
<td>246</td>
<td>896</td>
<td>1166</td>
<td>412</td>
<td>709</td>
<td>3429</td>
</tr>
</tbody>
</table>

Producer’s accuracy | User accuracy
-------------------|---------------
Mixed vegetation    | 53.3          | 33.0         | Overall accuracy = 57.34%  
Kappa = 0.50
Pine                | 36.0          | 44.2         
Open grassland       | 61.9          | 95.1         
Reeds                | 71.1          | 88.0         
Palmetto             | 70.2          | 41.1         

Table 2: accuracy assessment for the per-parcel classification

<table>
<thead>
<tr>
<th>Land-cover class</th>
<th>Forest vegetation</th>
<th>Pine</th>
<th>Open grassland</th>
<th>Reeds</th>
<th>Palmetto</th>
<th>Rows total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forest vegetation</td>
<td>49</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>50</td>
</tr>
<tr>
<td>Pine</td>
<td>17</td>
<td>144</td>
<td>11</td>
<td>0</td>
<td>42</td>
<td>214</td>
</tr>
<tr>
<td>Open grassland</td>
<td>2</td>
<td>6</td>
<td>365</td>
<td>0</td>
<td>29</td>
<td>402</td>
</tr>
<tr>
<td>Reeds</td>
<td>10</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>18</td>
</tr>
<tr>
<td>Palmetto</td>
<td>4</td>
<td>1</td>
<td>16</td>
<td>0</td>
<td>65</td>
<td>76</td>
</tr>
<tr>
<td>Columns total</td>
<td>82</td>
<td>152</td>
<td>385</td>
<td>4</td>
<td>137</td>
<td>760</td>
</tr>
</tbody>
</table>

Producer’s accuracy | User accuracy
-------------------|---------------
Mixed vegetation    | 59.8          | 98.0         | Overall accuracy = 82.4%  
Kappa = 0.73
Pine                | 94.7          | 67.3         
Open grassland       | 94.6          | 90.8         
Reeds                | 100.0         | 22.2         
Palmetto             | 47.4          | 85.5         
Table 3: accuracy assessment for the visually interpreted classification

<table>
<thead>
<tr>
<th>Land-cover class</th>
<th>mixed vegetation</th>
<th>Pine</th>
<th>Open grassland</th>
<th>Reeds</th>
<th>Palmetto</th>
<th>Rows total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixed vegetation</td>
<td>394</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>397</td>
</tr>
<tr>
<td>Pine</td>
<td>0</td>
<td>745</td>
<td>11</td>
<td>0</td>
<td>3</td>
<td>759</td>
</tr>
<tr>
<td>Open grassland</td>
<td>0</td>
<td>39</td>
<td>682</td>
<td>0</td>
<td>6</td>
<td>727</td>
</tr>
<tr>
<td>Reeds</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>333</td>
<td>0</td>
<td>333</td>
</tr>
<tr>
<td>Palmetto</td>
<td>0</td>
<td>6</td>
<td>64</td>
<td>4</td>
<td>1139</td>
<td>1213</td>
</tr>
<tr>
<td><strong>Columns total</strong></td>
<td><strong>394</strong></td>
<td><strong>790</strong></td>
<td><strong>760</strong></td>
<td><strong>337</strong></td>
<td><strong>1148</strong></td>
<td><strong>3429</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Producer's accuracy</th>
<th>User accuracy</th>
<th>Overall accuracy = 96.0%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixed vegetation</td>
<td>100</td>
<td>Mixed vegetation 99.2</td>
</tr>
<tr>
<td>Pine</td>
<td>94.3</td>
<td>Pine 99.6</td>
</tr>
<tr>
<td>Open grassland</td>
<td>89.7</td>
<td>Open grassland 93.8</td>
</tr>
<tr>
<td>Reeds</td>
<td>98.8</td>
<td>Reeds 100</td>
</tr>
<tr>
<td>Palmetto</td>
<td>99.2</td>
<td>Palmetto 94.0</td>
</tr>
</tbody>
</table>

4.2. Assessment of the variability of accuracy between different interpreters

Table 4 presents the results of the pixel-by-pixel comparison of the 30 classifications produced by different interpreters. Each interpreted map was compared in turn with the agreed ‘master’ classification. The numbers of pixels that were classified the same as the master were used to calculate the % correct for each class. The figures below are the average number of pixels correct from 30 different classifications.

<table>
<thead>
<tr>
<th>Ground reference dataset (GPS survey)</th>
<th>Mixed forest</th>
<th>Medium dense pine</th>
<th>Pine - Oak Mix</th>
<th>Open grassland</th>
<th>Reeds</th>
<th>Palmetto thicket</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>71.5 %</td>
<td>58.0 %</td>
<td>12.0 %</td>
<td>74.2 %</td>
<td>44.5 %</td>
<td>26.1 %</td>
</tr>
</tbody>
</table>

5. Discussion of results and conclusions

We explored the use of GPS ground survey to collect samples of ground reference data to estimate the accuracy of classifying IKONOS data of 1m spatial resolution using a variety of methods. We found that post-processing GPS observations using differential correction produced reference data of centimetre accuracy, which was more than sufficient to ensure precise matching of ground data to the corresponding image pixels. As well as increasing the positional accuracy requirement for the reference data, the use of high resolution (HR) data also increased the amount of ground reference data that needed to be collected for validating the classification. In this study, this was achieved by supplementing a data set of ground reference polygons collected previously for the purpose of validating medium resolution LANDSAT imagery, by collecting additional ground data within a more localised area covered by the IKONOS image. Given the increasing costs of ground surveys the number of image pixels for which ground reference data can be collected is likely to reduce, whilst the resolution of the imagery increases; this makes it more important to ensure the quality and representativeness of the reference data sample.

Using a common set of reference data obtained from a GPS ground survey, we see from tables 1-3 that the per-parcel classifier achieved higher accuracy for all the required land cover classes than the per pixel ML classifier. The overall accuracy for the per-parcel classifier was 82% (producer accuracy range: 47-95%, k=0.73) compared to 57.3%, (range: 36-70%, k=0.50) for ML. The classification by expert visual interpretation of a smaller extract within the image yielded an overall accuracy of 96% (range: 89-100%, k=0.95) when validated against the same ground reference data. These results suggest that for a typical mosaic of savanna land cover types, a per-parcel classifier can achieve higher accuracy than a per-pixel classifier (both overall and per-class). The per-parcel classification exceeded the minimum accuracy
requirement of 70% for the pine and grassland land cover classes and approached the target of 85% suggested for the overall accuracy required for natural resource mapping [15]. Whilst the per-parcel classification was not as accurate overall as a visual classification by an expert interpreter with local knowledge, we conclude that the per-parcel classifier can achieve an acceptable standard of accuracy for some of the savanna land cover classes, but that further work would be needed to improve the classification of smaller groups of trees and for sparser areas of woodland, before it could be used automatically. Some of the problems we experienced with using automated classifiers on higher resolution data are similar to those reported by [1], who noted that the automated classifiers were sometimes unable to delimit boundaries of features because the “within-class” variability was too high. They concluded that a combined approach using colour, texture and context was needed, which could be easier to achieve by visual interpretation.

Comparing a series of 30 manual, visual interpretations of the same image by practitioners in a developing country with differing degrees of experience of using HR optical data and different familiarity with the vegetation types, we see from table 4 that only the high forest and open grassland vegetation was classified with an average accuracy across all interpreters of above 70%. No single interpreter produced a map with an overall accuracy exceeding 85%. Even amongst those interpretations that were over 60% accurate, there was still considerable variability in where boundaries were drawn between the land cover classes. Whilst mixed forest and open grassland could be delimited reliably, interpreters often disagreed on the location and extent of pine and palmetto areas. A study by [16] also found that increasing the spatial resolution of imagery, while providing more descriptive information about patterns in the data, did not allow these patterns to be extracted reliably from image data.

The conclusion is that whilst a semi-automatic classification using a per parcel method may be less accurate than a visual classification by an expert interpreter, there can be significant variability in the areas delineated by different interpreters, even if these classifications achieve similar quantitative measures of class accuracy when checked against reference data. The goal of using a per-parcel classifier should perhaps be to consistently produce results of reasonable quality. This may be acceptable in cases where sufficient expert interpreters are not available, where it is not possible to check the consistency of the different interpretations or where mapping needs to be produced rapidly for extensive areas.

Although the two approaches of manual and per-parcel classifications may appear to be at either end of the spectrum of technological sophistication, in practice the two methods are often used together; users of sophisticated image processing software commonly assess how well a classification result conforms to, or is able to replicate a preliminary interpretation of the imagery ‘by eye’. This study has directly compared the accuracy of the two techniques while revealing some of the hidden uncertainty in visual classification results.

6. Acknowledgements

We extend special thanks to Alasdair MacArthur for supporting our use of E-cognition software, to Duncan Moss for his help processing the GPS survey observations and to Iain Woodhouse and Peter Furley for their assistance with running the image interpretation workshops in Belize. Funding for various fieldwork missions to collect the ground reference data was provided by the Royal Geographical Society, Royal Institute of Chartered Surveyors, Carnegie Trust for the Universities of Scotland, International Federation of Surveyors, Royal Scottish Geographical Society, Earth and Space Foundation, Edinburgh Earth Observatory and the University of Edinburgh. Funding for the image interpretation workshop was provided by a knowledge transfer grant from the Royal Scottish Geographical Society and the Scottish Executive. We are grateful to Drs Elma Kay and Ed Boles for kindly hosting this workshop at the University of Belize.

7. References


