Spatial analysis of remote sensing image classification accuracy

Alexis Comber [a,*], Peter Fisher [a], Chris Brunsdon [b], Abdulhakim Khmag [a]

[a] Department of Geography, University of Leicester, Leicester, LE1 7RH, UK
[b] Department of Geography, University of Liverpool, Liverpool, L69 3BX, UK

ARTICLE INFO

ABSTRACT

The error matrix is the most common way of expressing the accuracy of remote sensing image classifications, such as land cover. However, it and the measures that can be calculated from it have been criticised for not providing any indication of the spatial distribution of errors. Other research has identified the need for methods to analyse the spatial non-stationarity of error and to visualise the spatial variation in classification uncertainty. This research uses geographically weighted approaches to model the spatial variations in the accuracy of both (crisp) Boolean and (soft) fuzzy land cover classes. Remotely sensed data were classified using a maximum likelihood classifier and a fuzzy classifier to predict Boolean and fuzzy land cover classes respectively. Field data were collected at sub-pixel locations and used to generate soft and crisp validation data. A Geographically Weighted Regression was used to analyse spatial variations in the relationships between observations of Boolean land cover in the field and land cover classified from remote sensing imagery. A geographically weighted difference measure was used to analyse spatial variations in fuzzy land cover accuracy. Maps of the spatial distribution of accuracy were created for fuzzy and Boolean classes. This research demonstrates that data collected as part of a standard remote sensing validation exercise can be used to estimate mapped, spatial distributions of accuracy that would augment standard accuracy measures reported in the error matrix. It suggests that geographically weighted approaches, and the spatially explicit representations of accuracy they support, offer the opportunity to report land cover accuracy in a more informative way.

© 2012 Elsevier Inc. All rights reserved.

1. Introduction

Land cover information can be generated through the classification of remotely sensed data. Areas or pixels with similar spectral characteristics are allocated to classes or categories each of which represents a different type of land cover feature. It is a process of generalisation, and involves a number of choices about image type, resolution, number and types of classes, training sites, etc. (Campbell, 2007). Assessing map accuracy in an objective manner is fundamental to most land cover mapping projects (Foody, 2002; Strahler et al., 2006). The accepted paradigm for doing this is through comparison with some alternative data in order to determine measures of accuracy which "express the degree of 'correctness' of a map or classification" (Foody, 2002, p186). Determining the accuracy of land cover classified from remotely sensed data is important. Land cover is an input into environmental models incorporating land-atmosphere interactions (GLP, 2005) and land cover change is a major variable in climate change analyses (Feddema et al., 2005). In this context, accuracy descriptions can help the user to assess the uncertainties associated with incorporating land cover data into

their model or to decide between land cover datasets, especially where there is a choice between data with different thematic or spatial characteristics (See & Fritz, 2006). Thus accuracy is one of the key aspects of any remotely sensed data product.

The most common approach for assessing thematic map accuracy is to compare the classified land cover with alternative but spatially and temporally coincident data, which are considered to be of higher accuracy. A sample of the land cover data created by the remote sensing analysis (here referred to as 'classified' data) is compared against some validation data (here referred to as 'reference' data). The resulting cross tabulation of classified data against reference data is commonly known as the *error* matrix, but in the literature is also called the *confusion*, *contingency* or *validation* matrix. The cross tabulation provided by the error matrix allows a number of standard reporting measures to be calculated including overall accuracy as well as user's and producer's accuracies (Congalton, 1991; Congalton & Green, 1999). These accuracy statistics provide measures of the reliability of the classified data and the degree (but not spatial extent) to which they are correct. Therefore the appropriateness of the information conveyed by the error matrix may be limited when specific local conditions vary, for example when non-stationary error distributions occur, or in the presence of heteroscedastic residual distributions — i.e. when sub-sets of the data vary from the overall trend (Stehman, 2000, 2006).

There are two related limitations associated with accuracy assessments and error summaries calculated from the error matrix (McGwire & Fisher, 2001):

1) The error matrix and the accuracy measures it supports provide no information about the spatial distribution of error;
2) The overall accuracy measures derived from the error matrix may be inappropriate for sub-regions, where local error rates may be much larger or smaller than the global measures.

Overcoming such problems is important because many users of land cover data may be interested only in a particular subset of the data, either relating to a specific locale or to specific classes.

Some work in the remote sensing literature has explored the spatial distribution of different types of error and methods for reporting it. Campbell (1981) compared Landsat multispectral scanner images in the same growing season and found that misclassified pixels tended to be clustered. Congalton (1988) applied a Getis and Ord approach to analyse join count statistics to compare two datasets. Steele et al. (1998) used kriging to provide an optimal interpolation of map error. McGwire and Fisher (2001) recommended the use of Monte Carlo approaches to model the spatial distribution of errors. Some more recent research has examined the variability or non-stationarity of the distribution of errors. Riemann et al. (2010) describe a number of metrics for characterising the accuracy of spatial data that are dependent on reference data properties and Foody (2005) estimated local accuracy measures by interpolating the outputs of confusion matrices calculated at regular spaced intervals. Current validation and accuracy techniques in remote sensing have largely ignored the advances supported by such methods.

This research is in the spirit of Foody (2005). It uses Geographically Weighted Regression (GWR), a statistical method that explicitly deals with spatial non-stationarity (Brunsdon et al., 1996; Fotheringham et al., 2002), and a geographically weighted difference measure to analyse the spatial variations in the relationship between reference data and classified data for Boolean and fuzzy classes respectively. Geographically weighted approaches estimate spatially distributed measures of accuracy that are more informative than those provided by the confusion matrix. The paper proceeds as follows. Section 2 describes some of the scientific background to error matrices and their use in Boolean and fuzzy classifications. The methods and GWR are described in Section 3. Section 4 presents the results before a discussion of the issues arising from this research (Section 5) and some conclusions are drawn (Section 6).

2. Background

It is typical for the quality of spatial data such as land cover from remotely sensed imagery to be described using measures of thematic accuracy. The origins of the requirement of at least 85% thematic map accuracy can be traced back to Anderson (1971). Although the scientific basis for this accuracy level has been criticised (Congalton & Green, 1999; Pontius & Millones, 2011), it is historically related to land information being used for taxation assessments (Fisher, 1991; Fisher et al., 2002). Attribute accuracy is included in the major standards for spatial data quality reporting and metadata (Comber et al., 2008). Post classification comparisons of thematic data using the error matrix are the accepted method for describing the attribute accuracy of land cover classified from remotely sensed data.

One of the key issues associated with the error matrix relates to the use of crisp, Boolean data. Many of the key developments associated with confusion matrices and the measures of accuracy they support, are predicated on per pixel comparisons between reference and classified data. Analyses of crisp land cover classes are facilitated by the per pixel comparison embedded in the confusion matrix. It is becoming increasingly recognised within the remote sensing research community that soft classifications can provide a more representative model of the real world (Arnot & Fisher, 2007; Dronova et al., 2011; Fisher, 2010; Oldeland et al., 2011; Phillips et al., 2011; Rocchini, 2010). They allow some of the uncertainty associated with a pixel view of the world (Fisher, 1997) to be explicitly accommodated in the outputs of remote sensing analyses. In fuzzy classifications pixels can have partial memberships to different classes and the assumption of crisp membership embedded in the error matrix may be inappropriate. A number of extensions to the error matrix have been proposed by different authors, a selection of which is reviewed below.

Methods for assessing the accuracy of fuzzy classifications were suggested by Fisher and Pathirana (1991) based on estimating the portions of any pixel occupied by candidate land covers and were extended by Lewis and Brown (2001). Gopal and Woodcock (1994) suggested a fuzzy error matrix that incorporated linguistic descriptors to evaluate the land cover attributes associated with each sample location against each category in the classification scheme. Woodcock and Gopal (2000) illustrated the application of this method. The descriptors were used to calculate a set of discrete membership values to each map class at each sample site, which were then compared with fuzzy memberships extracted from remotely sensed data. Pontius and Cheuk (2006) developed this approach to describe an extended fuzzy error matrix that compared reference and classified fuzzy memberships. Their matrix contained multiple entries at each cell in the cross tabulation, describing the results of applying Boolean, Multiplication, Minimum and Composite operators. In particular they emphasised the applicability of the composite operator for multiple resolution analysis. More recently Lowry et al. (2008) extended the Gopal and Woodcock (1994) method and developed an approach that considered the off-diagonal elements in a fuzzy error matrix using a decision framework incorporating sets of rules, criteria and contexts within which class similarities are evaluated.

A second major reservation in relation to the error matrix is that is does not deal with geographic space very well. Foody (2005, p1218) says "The confusion matrix and the estimates of classification accuracy derived from it provide no information on the spatial distribution of error". The error matrix has been adopted as both the de facto and the de jure standard; *the* way to report on the accuracy of any remotely sensed data product. However, it does not allow the spatial distribution of errors to be described or reported, despite these having been found to be spatially autocorrelated in many studies over a long period of time. Some early examples include Campbell (1981), Labovitz (1984), Congalton (1988) and more recent examples can be found in Chen and Wei (2009) and Gonzalez et al. (2010). The implications of the spatial autocorrelation of errors are that different sub-regions of the study area will have different accuracies (e.g. Loveland et al., 1999; McGwire & Fisher, 2001) due to systematic errors associated with sensor, to the properties of the classes under investigation or because of the nature of the landscape, which the error matrix is unable to communicate. More recent research has considered the spatial information that *can* be obtained from the error matrix. Pontius and Millones (2011) show how the marginal totals can be used to estimate the quantity disagreement, whilst entries within the matrix are needed to compute allocation disagreement. Thus, although the error matrix and the typical measures that are calculated from it are aspatial and provide no information on the spatial distribution of error, some recent work has started to explore these issues in greater depth.

GWR has been used in some remote sensing research. Foody (2003) used GWR to analyse the relationship between normalized difference vegetation index (e.g. Richards, 1993), and rainfall and illustrated the extent to which global regression models provided poor local descriptions of the relationship. Similarly Wang et al. (2005) used GWR to examine the relationships between net primary production and a range of environmental variables. They noted that GWR made better predictions than ordinary least squares regression due to the spatial autocorrelation of the features under observation. Foody (2005) developed local accuracy descriptions using a moving

window to derive local estimates of thematic classification accuracy. Local confusion matrices were calculated from the nearest 150 out of 1000 sample points at 48 locations evenly distributed across the study area. Accuracy surfaces were estimated from these locations using an inverse distance squared interpolation algorithm, reporting for example the user's accuracy for winter wheat and so on. Three interesting and salient aspects of Foody's research were the use of a large number of sample points in a small area (and Foody notes that local accuracy assessment may not be appropriate in circumstances where the number of sample points was low), the use of a fixed window (compared to the dynamically optimised one described below), and second the lack of any distance weighting — all of the 150 points contributed equally to the result at each location.

This research builds on Foody's, 2005 work and extends the use of explicitly geographical approaches to remote sensing accuracy assessment. It addresses two long-standing gaps in the analysis and communication of error in thematic land cover data. First, it uses geographically *weighted* methods to describe the varying spatial relationship between reference and classified land cover classes to for Boolean and fuzzy class assignments. Second, it uses these methods to determine the non-stationarity of errors to show how they vary discontinuously (McGwire & Fisher, 2001) and to visualise spatial variations in classification errors (Foody, 2002).

3. Methods

3.1. Data and study area

The area of the present study is located in the north western part of Libya in Jifara Plain, around Tripoli. Satellite imagery from the Système Pour l'Observation de la Terre (SPOT) 5 sensor from 2009 was resampled to 30 m × 30 m as part of a wider study examining land cover changes using Landsat data from 1976, 1989 and 2005. It was classified into 6 classes: Urban, Woodland, Vegetation, Grazing Land and Bare areas and Water. Water was not a focus for this research. The class descriptions are shown in Table 1. Two supervised classifications were generated: a standard, crisp Boolean classification using the maximum likelihood classifier in Idrisi (the *Maxlike* module), and a fuzzy classification model using the fuzzy *c*-Means described in Wang (1990) as implemented in the *Fuzclass* module in Idrisi. Fuzzy *c*-Means defines fuzzy class memberships based on distance to the class centroid as defined by the mean of all points, weighted by their degree of belonging to the class. Both classifications used the same training data.

In classic set theory an object can be assigned values of only 0 (non-membership) or 1 (membership) to each element *x* of a universal set *X* to indicate whether that object belongs to a certain crisp, Boolean set A. In Fuzzy Set theory this is generalized into the membership function, which allows for the assignment of a value in the range [0,1] to each element *x* of the universal set, corresponding to grades of

membership of element *x* to the Fuzzy Set A (Klir & Yuan, 1995). The use of Fuzzy Sets in remote sensing is described by Fisher and Pathirana (1991) and by Foody (1996). This approach calculates a degree of membership to each class for each pixel in the range [0,1] as defined by the training data and a similarity relation model.

A validation dataset was collected by field survey. The study area was divided into 21 blocks of approximately 100 km^2 and 10 sample locations in each block were selected randomly. Thus, sampling locations were identified using a spatially stratified random scheme. Each sample location represented a pixel. At each sample location the overall land cover class and the land cover at 16 points in a 4 × 4 grid within a 30 m × 30 m area were recorded. Precise sub-pixel locations were established using differential GPS. The sub-pixel measures of land cover were combined to create fuzzy memberships to the different land cover classes for each pixel location. Consequently the fuzzy data collected in the field at the validation locations were limited to only one of 17 values between 0 and 1 (0/16, 1/16, 2/16, 3/16 …16/16), whereas the fuzzy classification of remotely sensed data were allowed to take any value between 0 and 1. The reference data points and their locations are shown in Fig. 1. The spatial distribution of the Boolean data, the fuzzy memberships as determined from the remotely sensed data and as observed in the field are shown in Fig. 2 for each land cover class.

For the analysis of the Boolean data, a geographically weighted logistic regression analysed binary data indicating the presence (1) or absence (0) of a particular land cover class for both reference (validation) and classified land cover. For the analysis of the fuzzy data, a geographically weighted difference analysis compared fuzzy membership values in the range [0,1] of the fuzzy reference and the classified land cover. In the geographically weighted analyses described below, spatially distributed estimates of the coefficients (Boolean) and of the mean differences (fuzzy) were calculated to describe the relationships between classified and reference validation data.

3.2. Geographical analysis

Geographically weighted approaches were used to analyse the spatial variations in the relationships between the reference data collected in the field and classified data from the remote sensing analysis. A 'geographically weighted' approach is one that uses a moving window or kernel and applies a distance weighting to the data. A Geographically Weighted Regression computes local estimates of the regression coefficients from the distance weighted data under the kernel at locations throughout the study area. Linear regression is used in many areas of science and implicitly assumes spatial stationarity of the relationships identified in the analysis — it assumes that the relationships between the variables remain constant over geographical space. When a relationship (or pattern) that applies in one region does not apply in another, spatial non-stationarity is said to occur. GWR accommodates the possibility that relationships vary over geographical space by allowing regression coefficient estimates to vary with location and the relative likelihood of a correct prediction to change spatially. A full description of GWR can be found in Fotheringham et al. (2002) and Brunsdon et al. (1996), but the generic process is summarised here.

3.2.1. Boolean accuracy

In overview, the approach was to logistically regress the reference Boolean class (dependent variable) on the classified Boolean class (independent variable). This was carried out for each class.

The *logit* function is defined by:

$$\mathrm{logit}(Q) = \frac{\exp(Q)}{1 + \exp(Q)} \tag{1}$$

where Q is any value.

Table 1
Land cover class descriptions.

| Classes	Descriptions of land cover classes
Urban	Areas characterised by buildings, asphalt, concrete, and artificial surfaces.
Woodland	Land dominated by trees, including natural woodland and woody plantations.
Grazing Land	Areas where herbivores feed on plants; any vegetated land such as small shrubs that is grazed or that has the potential to be grazed by animals.
Vegetation	All other types of vegetation, including shrubs, and crops, whether irrigated or rainfed.
Bare area	Areas of bare rock, sand, silt, gravel or other earthen material with little or no vegetation including beaches and sandy areas.
Water	All types of water such as sea and lakes.
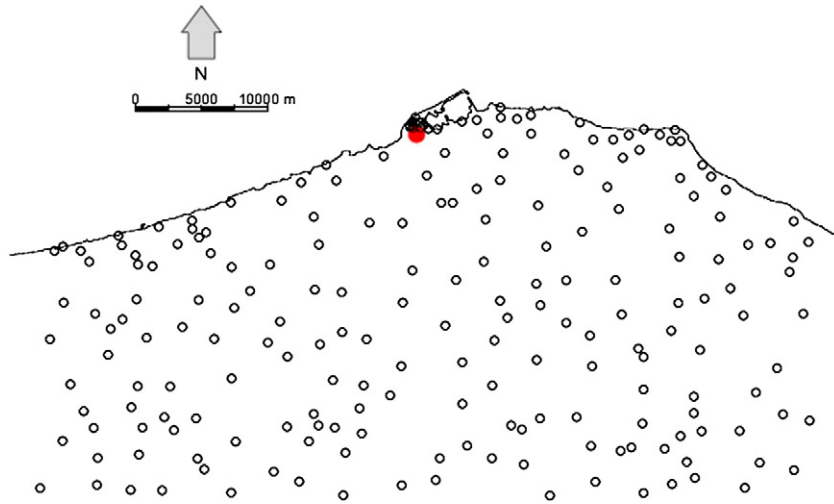
Fig. 1. The study area in Libya, around the port of Tripoli (red circle in the North), and the location of the 210 reference data sites.

Fig. 2. The data collected in the field and classified from remotely sensed (RS) data. In the Boolean maps, solid symbols (●) indicate where both Field and RS agree, hollow circles (○) where only the RS indicates the class and crosses (+) where only field survey indicates the class. In the maps of fuzzy classes, the size of the plot characters indicates the degree of membership to the class.

The logistic Geographically Weighted Regressions were calculated as follows:

$$pr(y_i = 1) = \text{logit}\left(b_{0(u_i,v_i)} + b_1 x_{1(u_i,v_i)}\right) \quad (2)$$

where $pr(y_i = 1)$ is the probability that the reference Boolean class is present, x_{1i} is the explanatory or independent variable (the presence of the classified Boolean class) and the coefficient estimates for the explanatory variable are assumed to vary across the two-dimensional geographical space defined by the coordinates (u, v). Thus the coefficient estimates in the logistic GWR can be considered as functions of these coordinates, rather than constants.

In contrast to global models where processes are assumed to be stationary (i.e. location invariant), local models are spatial disaggregations of global models, the results of which are location-specific. So, the template of GWR is similar to an ordinary logistic regression model but the coefficient estimates are allowed to vary geographically using a kernel function. A moving window allows a local regression analysis to be computed at each location with points that are further away from the specific location under consideration contributing less to the solution. That is, a weighted regression is carried out where the weight, w_i, associated with each location (u_i,v_i) is a decreasing function of d_i, the distance from the centre of the window to (u_i,v_i):

$$w_i = \begin{cases} \left(1 - \dfrac{d_i^2}{h^2}\right)^2 & \text{if } d_i < h \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where h is known as the *bandwidth* of the GWR calibration. In this way the weights associated with each location change depending on the location for which a prediction is to be calculated.

The kernel function defines the data and weights that are used to calibrate the model at each location. The bandwidth may be varied to ensure that enough data points are used in the calibration to minimise the cross validation prediction error. Usually the number of data points is a trade-off between working with a dataset that is too small to calibrate the local model reliably and too big to avoid averaging out local effects. In this case, the bandwidth was selected automatically by leave-one-out cross validation, which finds a bandwidth for a given geographically weighted analysis by optimising the ability of the GWR (Eq. 2) and geographically weighted difference (Eq. 5 below) analyses to predict each individual reference value when it has been removed from the dataset. Further details on bandwidth selection can be found in Fotheringham et al. (2002). The bandwidths for the Boolean and fuzzy analyses were adaptively defined to determine the number of data points required to minimise the cross validation prediction error. For example, the Boolean class of Urban required a bandwidth that included 9.0% of the total number of data points and the fuzzy class of Grazing Land a bandwidth that incorporated 6.2% of them. Thus, by allowing bandwidths to vary, by using spatially dependent weights and by computing estimates of the regression coefficients locally, GWR is an approach that addresses spatial non-stationarity. It reflects one of the fundamental tenets of geographical analyses by operating under the assumption that the effect of the explanatory variables on the dependent variable will vary continuously over space.

The probabilities of correctly identifying Boolean reference class, y, given the classified data, x, were estimated as follows:

$$pr(y = x) = pr(y = 1|x = 1)pr(x = 1) + pr(y = 0|x = 0)pr(x = 0) \quad (4)$$

where a value of 0 denotes the absence of that class and 1 denotes the presence for both x and y, and $pr(statement\ 1|statement\ 2)$ denotes the probability that *statement 1* is true given that *statement 2* is true. Therefore the right hand side of Eq. (4) denotes the probability that the reference and classified Boolean statements agree. In this case the terms $pr(y=1|x=1)$ and $pr(y=0|x=0)$ were estimated using

the values obtained in Eq. (2). The terms $pr(x=0)$, the proportion of cases for which x = 0, and $pr(x=1)$, the proportion of cases for which x = 1, were determined from the Boolean data from the counts of the number of times the class was either 0 or 1. Note that because a geographically weighted approach was used, a different value for $pr(y=x)$ is associated with each location (u, v).

This is a portmanteau measure of accuracy that includes both specificity and sensitivity — i.e. the probability that that either the presence or absence of a particular land cover class is correctly classified. This is further illustrated by consideration of Table 2, where n_1, n_2, n_3 and n_4 represent the full set of probabilities at each location, as estimated by the geographically weighted approach. Whilst user's accuracy is $n_1/(n_1 + n_2)$ and producer's accuracy is $n_1/(n_1 + n_3)$, in this case we are measuring portmanteau accuracy using $(n_1 + n_4)/(n_1 + n_2 + n_3 + n_4)$. Thus portmanteau accuracy describes the overall accuracy when the data are collapsed to two classes, the land cover type of interest, and all other land cover types combined into a single class. Although global measures of portmanteau accuracy could be computed without using logistic regression, specifying the model using GWR allows spatial variations in accuracy to be identified.

3.2.2. Fuzzy accuracy

A fuzzy geographically weighted difference analysis for each land cover class was applied under the following logic: y is the fuzzy membership of the reference data based on observation in the field and x is the fuzzy membership of the classified data from the remote sensing analysis. Now consider classified and reference fuzzy membership values to a given class at a given location: the difference measure is defined to be the absolute difference, D between the two fuzzy memberships:

$$D_{(u,v)} = abs\left(y_{(u,v)} - x_{(u,v)}\right) \quad (5)$$

where y_i is the reference fuzzy membership, x_{1i} is the classified fuzzy membership and (u, v) the two-dimensional geographical space. Note that if all membership functions were Boolean, such that y and x can only take values 1 and 0, then D also only takes values 1 and 0, and that if these are interpreted as *true* and *false* respectively, then this corresponds to the membership of the Boolean set of locations where the two classifications disagree. Thus, the quantity D defined above can be thought of as a fuzzy generalisation of the Boolean situation described in Section 3.2.1. As D can take any value in the range 0 to 1, with 0 corresponding to a crisp classification of agreement and 1 corresponding to a crisp classification of disagreement between reference and classified land classes, it can be thought of as a fuzzy measure of accuracy.

Having defined this measure of accuracy, it is now necessary to consider spatial patterns in the levels of accuracy. One approach is to compute a surface of weighted moving window means of D, whose weights use the same kernel as defined in Eq. (3) in a similar manner to GWR — see for example Brunsdon et al. (2002). This surface can then be used to visualise spatial trends in the fuzzy accuracy measure.

All of the statistical analysis and mapping were implemented in R version 2.15.1, the open source statistical software http://cran.r-project.org. The *spgwr* library was used to perform the geographically weighted analyses. The maps were created with the *GISTools* library. The data and code developed for this analysis will be provided to interested researchers on request.

Table 2
The 2-class error matrix suggested by the probabilities in Eq. (4).

	Reference true (y = 1)	Reference false (y = 0)
Classified true (x = 1)	n_1	n_2
Classified false (x = 0)	n_3	n_4

Table 3
Error matrix comparing reference and classified Boolean land cover.

Classified	Reference					Total	User
	Bare	Grazing Land	Urban	Vegetation	Woodland		
Bare	18	8	7	2	4	39	0.46
Grazing Land	3	23	3	8	6	43	0.53
Urban	0	0	27	1	2	30	0.90
Vegetation	0	4	7	31	5	47	0.66
Woodland	0	4	2	18	27	51	0.53
Total	21	39	46	60	44		0.60
Producer	0.86	0.59	0.59	0.52	0.61	0.60	Overall

Table 4
A summary of the variation portmanteau accuracies — i.e. the probability that the presence or absence of each class is correctly predicted in the reference data.

Class	Min.	1stQu.	Median	Mean	3rdQu.	Max.	IQR
Bare	0.883	0.883	0.885	0.885	0.887	0.888	0.004
Grazing Land	0.741	0.803	0.819	0.829	0.864	0.931	**0.060**
Urban	0.776	0.852	0.905	0.898	0.937	0.996	**0.086**
Vegetation	0.785	0.787	0.787	0.788	0.788	0.790	0.002
Woodland	0.800	0.802	0.809	0.807	0.811	0.812	0.009

4. Results

The validation data were used to construct a standard error matrix (Table 3). The crisp, Boolean data allow a straightforward comparison between data classified from the remote sensing imagery and the reference data collected in the field, as well as user's and producer's accuracies to be calculated. Whilst it is evident that some classes are more reliably classified than others, the table provides no information about the spatial distribution of either overall error or errors for different classes.

4.1. Boolean accuracy

A geographically weighted logistic regression was used to explore the spatial variation in the relationship between reference and classified Boolean data. Summaries of the portmanteau accuracy of Boolean classes using Eq. (4) are shown in Table 4, where the minimum, median, maximum and 1st and 3rd quartiles of the probabilities of the class predictions being correct are reported, together with the global

Table 5
A summary of the variation of the mean differences between reference and classified fuzzy classes.

Class	Min.	1stQu.	Median	3rdQu.	Max.	Global	IQR
Bare	0.158	0.158	0.159	0.160	0.160	0.159	0.002
Grazing Land	0.058	0.137	0.192	0.237	0.304	0.178	**0.100**
Urban	0.022	0.057	0.089	0.141	0.299	0.110	**0.084**
Vegetation	0.173	0.179	0.192	0.203	0.211	0.191	0.024
Woodland	0.161	0.163	0.163	0.164	0.165	0.164	0.001

value and the inter-quartile range (IQR). These show two things: the probabilities that the reference data are correctly predicted by the Boolean remote sensing classification and the variation of these probabilities. The IQR gives an indication of the overall spatial variation in these accuracies. Table 4 shows that there are greater spatial variations in the relationships between classified and reference data for the Urban and Grazing Land classes highlighted in bold. The mappings in Fig. 3 illustrate these variations.

4.2. Fuzzy accuracy

A geographically weighted difference analysis was used to examine the spatial variation in the difference between reference and classified fuzzy memberships. Table 5 shows the variations in these differences for each class. The low numbers in Table 5 indicate that the differences between the classified and reference fuzzy classes are quite small, and similar to the Boolean results the greatest variation was found for the classes of Urban and Grazing Land, highlighted in bold. By subtracting the differences from unity, the accuracies were determined and Fig. 4 shows the spatial variations in accuracy.

4.3. Summary of results

The spatial distributions of Boolean portmanteau accuracy and fuzzy difference for the Urban and Grazing Land classes in Figs. 3 and 4 are closely linked to the spatial pattern of those land cover classes in the reference and classified data. Consider the situation where a class is absent over a part of the study area (i.e. no membership to the fuzzy or Boolean set of either the classified or the reference data), then the 'error' of the classified data relative to the reference data (either probability or difference) would be low. This can be seen to the south for the Urban class and to the North for the Grazing Land class in Fig. 5.
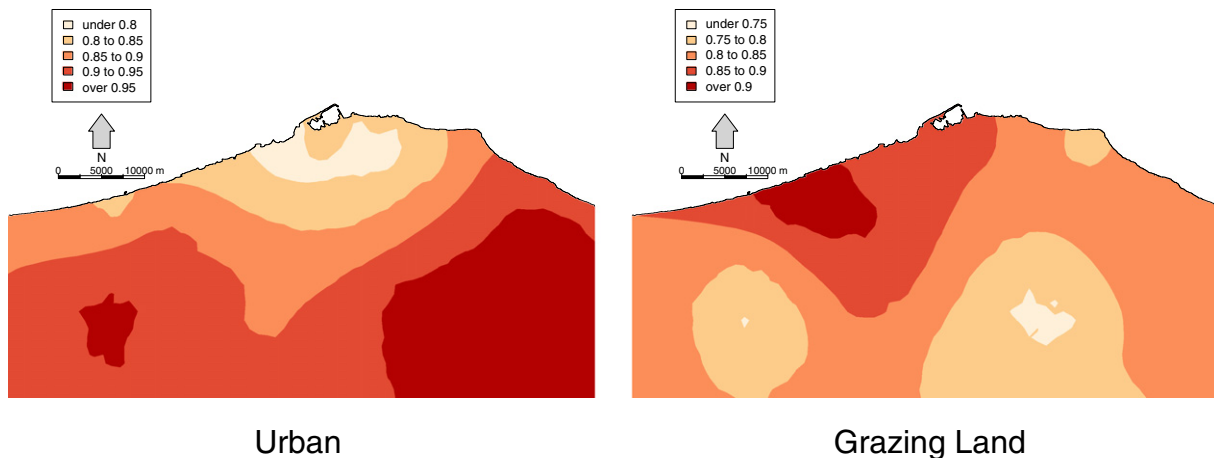

Urban　　　　　　　　　　　Grazing Land

Fig. 3. The spatial distribution of portmanteau accuracies for Urban and Grazing Land.
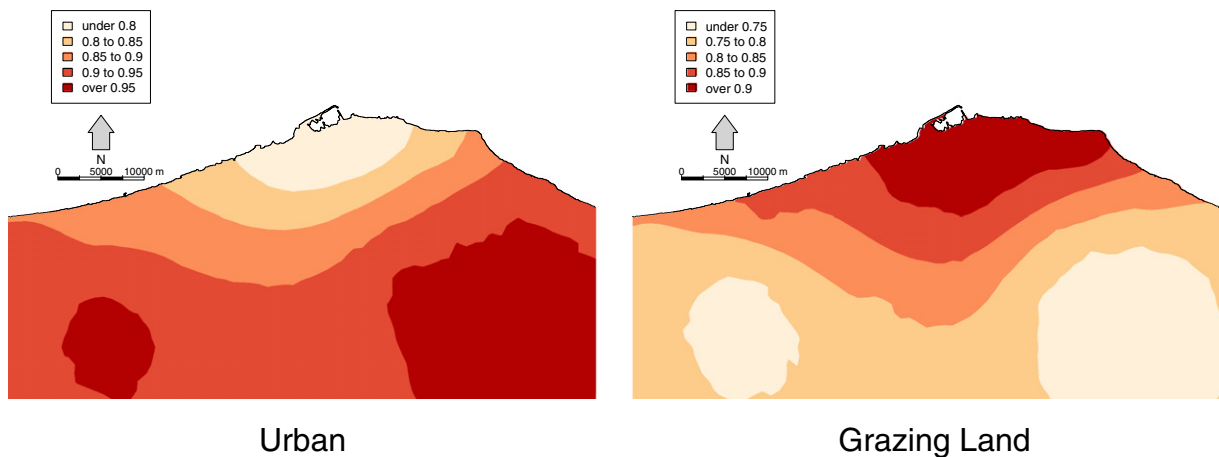
Fig. 4. The spatial distributions of the accuracies of fuzzy Urban and Grazing Land.

5. Discussion

The major contributions of this research relate to the development of 1) spatially distributed measures of accuracy using a kernel and distance weighting in geographically weighted accuracy measures; 2) a portmanteau accuracy measure for Boolean land cover data; and 3) a fuzzy difference measure to describe the accuracy of fuzzy classifications. The outputs of the Boolean analysis, using a portmanteau measure of accuracy, indicate the spatial variation in the extent to which the reference classes in the field were predicted by the remote sensing analysis. They describe the spatial variations in the probability of correctly identifying reference field values, given the classified data. The outputs of the fuzzy analysis, using a geographically weighted difference measure, provide spatially distributed measures of fuzzy prediction accuracy. Figs. 3 and 4 show broadly similar patterns for Boolean and fuzzy accuracy for the class of Urban and of Grazing Land, but with subtle local differences in each case, reflecting the different classification approaches.

The geographically weighted approach uses a kernel, defined in Eq. (3), in a similar way to Foody (2005) but with a number of critical differences. Foody analysed a dataset with 1000 validation points, compared to the 210 used in this study. From this, he computed local confusion matrices at 48 locations regularly distributed in the study area. Thus Foody's method arbitrarily partitions the study area into a fixed set of equal-sized sub-regions, computes accuracy for each sub-region, and then spatially interpolates over those sub-regions. Foody's analysis was 'geographical' in that he used the nearest 150 validation points to calculate the local confusion matrices and user and producer accuracy, but the effect of the distance of the validation point from the centre of the window was not taken into account. Whereas the method proposed in this study is 'geographically weighted' as it allows for the fact that the more distant observations may be of less local relevance. This means that in the presence of trends, where more peripheral points may reflect different levels of accuracy, their influence is weighted accordingly. In this way the application of geographically weighted methods in this research extends Foody's approach and reflects his observation that information on local accuracy can be calculated "at no extra cost to the analysis other than a small amount of time and effort" (p. 1226) from standard validation data.

Geographically explicit (i.e. local) analyses of the *spatial accuracy* of land cover classified from remotely sensed data have not been widely reported, despite books and conferences of this name originating from the remote sensing community, as well as research in other areas that have reported the spatial distribution of errors (Oksanen and Sarjakoski (2007) considered DEMs, for example). Evidence for this statement can be found in a review of the scientific literature of recent research and in the standards for spatial data quality reporting, where spatially static protocols for data quality metadata persist in all the major standards (OGC, ISO and Dublin Core). In part this situation may be driven by a lack of demand for spatially more informative descriptions of error and accuracy and it may also be that such developments have not been offered as a possibility by methodology researchers. The error matrix is the accepted paradigm for reporting the accuracy of thematic maps and a number of observations relating to the aspatial nature of the error matrix were made earlier in this paper. The static accuracy measures supported by the error matrix reflect a cartographic legacy of paper map production rather than digital mapping (Fisher, 1998). However, historically in cartography it was also a common practice to include small inset maps of reliability to report accuracy related information. This is particularly evident in thematic soil mapping. One example is the FAO/UNESCO Soil Map of the World (e.g. Volume V1–2 Africa), where small inset maps show whether information originates from "Systematic soil survey", "Soil reconnaissance" or from "General Information with local soil observations". The most reliable information is from the first of these and the last produces the least reliable. The opportunities offered by such paradigms have been ignored in the digital revolution: they may no longer be perceived to be relevant to digital products or to the way that they are used and treated. In part this may be related to the way that land cover products are commissioned: many land cover mapping projects are required to specify a minimum level of overall error, with an implicit assumption that it is evenly distributed. The methods described in this article estimates spatial distributions of accuracy and these offer the opportunity for more nuanced specifications of accuracy, for example a proportion of the area within ranges of acceptable error.

This research does not seek to overcome all of the shortcomings in the error matrix. Rather, this research explored spatially explicit methods for reporting accuracy using geographically weighted statistical methods to identify spatial variations in the relationship between classified and reference data. The maps in Figs. 3–5 provide spatial information on the distribution and variation in accuracy. They are not accompanied by standard errors, which may be high if a fixed kernel size is used (it was not in this research) because the accuracy at any given location is calculated from the sample observations defined by the kernel. Under a fixed kernel the number of sample observations may be small. However, the ability to estimate spatially explicit measures of accuracy and error from data collected as part of validation exercise, suggests that maps of the distribution of accuracy could accompany confusion matrices.

The spatially explicit outputs that can be produced using geographically weighted approaches indicate the potential advantages of including the results of any validation exercise along with the
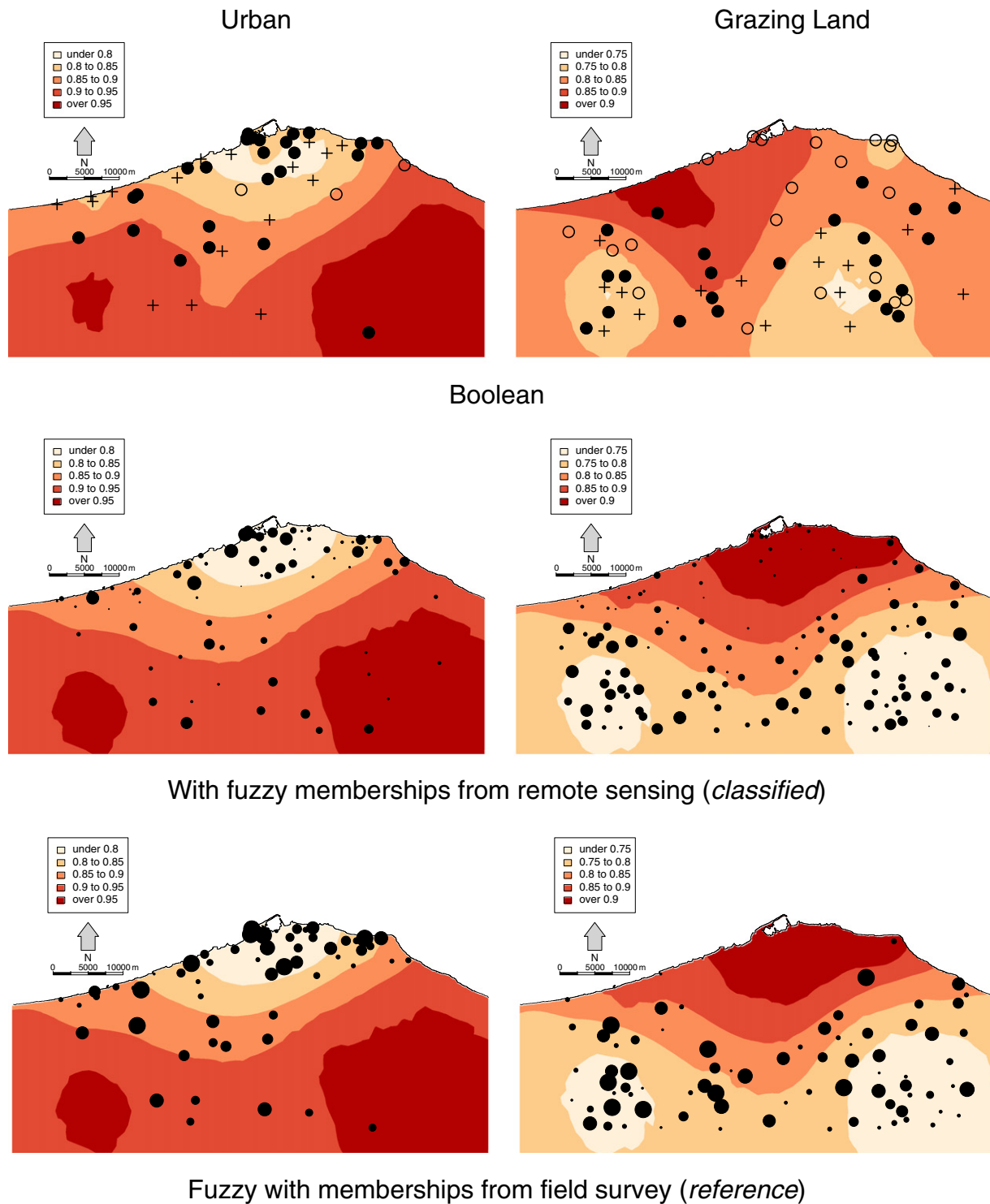
Fig. 5. The spatial variation of accuracies as in Figs. 3 and 4, but with the classified and reference data points. In the Boolean maps, solid symbols (●) indicate where they both agree, hollow circles (○) where only the data classified by remote sensing indicates the class and crosses (+) where only the reference data from the field survey indicates the class. In the maps of fuzzy classes, the size of the plot characters indicates the degree of membership to the class.

land cover data product. McGwire and Fisher (2001) commented that "the expansion of error documentation methods may include the addition of a spatial error summary in the product metadata, including at the least both the coordinates and the attribute data that were used in the standard product validation" (p328). Such practice, in combination with the methods described in this paper, would allow data users to explore the spatial distribution of errors and the

relationships between classes relative to their intended uses of the data. It would provide opportunities for users to apply dynamic analyses of such metadata as suggested by Comber et al. (2008). Finally, the methods introduced in this paper and by Foody (2005) support a potential methodological shift in the way that land cover error and validation are reported away from static, aspatial measures of accuracy, towards ones which are geographically explicit. Portmanteau

accuracies provide an alternative to user/producer accuracies, and the spatial variation in absolute fuzzy difference obviates the need for fuzzy confusion matrices.

6. Conclusions

This research uses geographically weighted approaches to describe the spatial variation in the accuracy of Boolean and fuzzy classifications of remotely sensed data. It proposes a portmanteau approach to describe Boolean land cover accuracy and fuzzy difference measures to describe the accuracy of fuzzy land cover. It addresses two long-standing gaps in the analysis and communication of accuracy and error land cover. First, by analysing the spatial distribution of errors it provides a better understanding of non-stationarity in land cover errors, which frequently vary in a discontinuous manner, than can be determined from the confusion matrix. Second, it uses a geographically weighted approach to estimate the spatial variation in classification accuracy, which can be mapped as accompaniments to the classification. Additionally, generic geographically weighted approaches – computing local measures under a kernel with a distance weighting – could be used for any statistic or calculation. Such approaches would better reflect Tobler's 1st law of geography: "Everything is related to everything else, but near things are more related than distant things" (Tobler, 1970, p236). Finally, the geographically weighted approaches using a moving window or kernel described in this research suggest potential advancements in the way that accuracy and error in spatial datasets are described. Spatially explicit measures of accuracy are more informative precisely because they are spatial and provide greater support to user assessments of data accuracy than the global measures derived from the confusion matrix. The research presented in this paper suggests that perhaps it is time to reconsider current thinking around the reporting of accuracy and error associated with remote sensing.

Acknowledgements

References

Anderson, J. R. (1971). Land use classification schemes used in selected recent geographic applications of remote sensing. *Photogrammetric Engineering*, *37*(4), 379–387.

Arnot, C., & Fisher, P. (2007). Mapping the ecotone with Fuzzy sets. In A. Morris, & S. Kokhan (Eds.), *Geographic uncertainty in environmental security* (pp. 19–32). Dordrecht: Springer.

Brunsdon, C. F., Fotheringham, A. S., & Charlton, M. (1996). Geographically weighted regression — A method for exploring spatial non-stationarity. *Geographical Analysis*, *28*, 281–298.

Brunsdon, C., Fotheringham, A. S., & Charlton, M. (2002). Geographically weighted summary statistics — A framework for localized exploratory data analysis. *Computers Environment and Urban Systems*, *26*, 501–524.

Campbell, J. (1981). Spatial correlation effects upon accuracy of supervised classification of land cover. *Photogrammetric Engineering and Remote Sensing*, *47*, 355–364.

Campbell, J. B. (2007). *Introduction to remote sensing* (4th ed.). London: Taylor and Francis.

Chen, D. M., & Wei, H. (2009). The effect of spatial autocorrelation and class proportion on the accuracy measures from different sampling designs. *ISPRS Journal of Photogrammetry and Remote Sensing*, *64*(2), 140–150.

Comber, A. J., Fisher, P. F., & Wadsworth, R. A. (2008). Using semantics to clarify the conceptual confusion between land cover and land use: The example of 'forest'. *Journal of Land use Science*, *3*(2–3), 185–198.

Congalton, R. G. (1988). Using spatial auto-correlation analysis to explore the errors in maps generated from remotely sensed data. *Photogrammetric Engineering and Remote Sensing*, *54*, 587–592.

Congalton, R. G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, *37*(1), 35–46.

Congalton, R. G., & Green, K. (1999). *Assessing the accuracy of remotely sensed data: Principles and practices.* Boca Raton: Lewis Publishers.

Dronova, I., Gong, P., & Wang, L. (2011). Object-based analysis and change detection of major wetland cover types and their classification uncertainty during the low water period at Poyang Lake, China. *Remote Sensing of Environment*, *115*(12), 3220–3236.

Feddema, J. J., Oleson, K. W., Bonan, G. B., Mearns, L. O., Buja, L. E., Meehl, G. A., et al. (2005). The importance of land-cover change in simulating future climates. *Science*, *310*, 1674–1678.

Fisher, P. F. (1991). Modelling soil map-unit inclusions by Monte Carlo simulation. *International Journal of Geographical Information Systems*, *5*(2), 193–208.

Fisher, P. F. (1997). The pixel: A snare and a delusion. *International Journal of Remote Sensing*, *18*(3), 679–685.

Fisher, P. F. (1998). Improved modelling of elevation error with geostatistics. *GeoInformatica*, *2*, 215–233.

Fisher, P. F. (2010). Remote sensing of land cover classes as type 2 Fuzzy sets. *Remote Sensing of Environment*, *114*, 309–321.

Fisher, P. F., Comber, A. J., & Wadsworth, R. A. (2002). The production of uncertainty in spatial information: the case of land cover mapping. In G. Hunter, & K. Lowell (Eds.), *Accuracy 2002, 5th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences* (pp. 60–65). Melbourne: The University of Melbourne.

Fisher, P. F., & Pathirana, S. (1991). The evaluation of Fuzzy membership of land cover classes in the sub urban zone. *Remote Sensing of Environment*, *34*, 121–132.

Foody, G. M. (1996). Approaches for the production and evaluation of Fuzzy land cover classification from remotely-sensed data. *International Journal of Remote Sensing*, *17*, 1317–1340.

Foody, G. M. (2002). Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, *80*, 185–201.

Foody, G. M. (2003). Geographical weighting as a further refinement to regression modelling: An example focused on the NDVI–rainfall relationship. *Remote Sensing of Environment*, *88*, 283–293.

Foody, G. M. (2005). Local characterization of thematic classification accuracy through spatially constrained confusion matrices. *International Journal of Remote Sensing*, *26*, 1217–1228.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. E. (2002). *Geographically weighted regression: The analysis of spatially varying relationships.* Chichester: Wiley.

GLP (2005). Science plan and implementation strategy. *IGBP Report No. 53/IHDP Report No. 19*. Stockholm: IGBP Secretariat.

Gonzalez, P., Asner, G. P., Battles, J. J., Lefsky, M. A., Waring, K. M., & Palace, M. (2010). Forest carbon densities and uncertainties from Lidar, QuickBird, and field measurements in California. *Remote Sensing of Environment*, *114*(7), 1561–1575.

Gopal, S., & Woodcock, C. (1994). Theory and methods for accuracy assessment of thematic maps using Fuzzy sets. *Photogrammetric Engineering and Remote Sensing*, *60*(2), 181–188.

Klir, G., & Yuan, B. (1995). *Fuzzy sets and fuzzy logic—Theory and applications.* New Jersey: Prentice Hall PTR.

Labovitz, M. L. (1984). The influence of autocorrelation in signature extraction — An example form a geobotanical investigation of Cotter basin, Montana. *Photogrammetric Engineering and Remote Sensing*, *52*, 201–211.

Lewis, H. G., & Brown, M. (2001). A generalized error matrix for assessing area estimates from remotely sensed data. *International Journal of Remote Sensing*, *22*(16), 3223–3235.

Loveland, T. R., Zhu, Z., Ohlen, D. O., Brown, J. F., Reed, B. C., & Yang, L. (1999). An analysis of the IGBP global land-cover characterization process. *Photogrammetric Engineering and Remote Sensing*, *65*(9), 1021–1032.

Lowry, J. H., Ramsey, R. D., Stoner, L. L., Kirby, J., & Schulz, K. (2008). An ecological framework for evaluating map errors using fuzzy sets. *Photogrammetric Engineering and Remote Sensing*, *74*(12), 1509–1519.

McGwire, K. C., & Fisher, P. (2001). Spatially variable thematic accuracy: Beyond the confusion matrix. In C. T. Hunsaker, M. F. Goodchild, M. A. Friedl, & T. J. Case (Eds.), *Spatial uncertainty in ecology: Implications for remote sensing and GIS applications* (pp. 308–329). New York: Springer–Verlag.

Oksanen, J., & Sarjakoski, T. (2007). Uncovering the statistical and spatial characteristics of fine toposcale DEM error. *International Journal of Geographical Information Science*, *20*(4), 345–369.

Oldeland, J., Dorigo, W., Lieckfeld, L., Lucieer, A., & Jürgens, N. (2011). Combining vegetation indices, constrained ordination and Fuzzy classification for mapping semi-natural vegetation units from hyperspectral imagery. *Remote Sensing of Environment*, *114*(6), 1155–1166.

Phillips, T., Leyk, S., Rajaram, H., Colgan, W., Abdalati, W., McGrath, D., et al. (2011). Modeling moulin distribution on Sermeq Avannarleq glacier using ASTER and WorldView imagery and Fuzzy set theory. *Remote Sensing of Environment*, *115*(9), 2292–2301.

Pontius, R. G., Jr., & Cheuk, M. L. (2006). A generalized cross-tabulation matrix to compare soft-classified maps at multiple resolutions. *International Journal of Geographical Information Science*, *20*, 1–30.

Pontius, R. G., & Millones, M. (2011). Death to Kappa: Birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, *32*(15), 4407–4429.

Richards, J. A. (1993). *Remote sensing digital image analysis: An introduction, 340p.* Berlin: Springer.

Riemann, R., Wilson, B. T., Lister, A., & Parks, S. (2010). An effective assessment protocol for continuous geospatial datasets of forest characteristics using USFS Forest Inventory and Analysis (FIA) data. *Remote Sensing of Environment*, *114*, 2337–2352.

Rocchini, D. (2010). While Boolean sets non-gently rip: A theoretical framework on fuzzy sets for mapping landscape patterns. *Ecological Complexity*, *7*, 125–129.

See, L. M., & Fritz, S. (2006). A method to compare and improve land cover datasets: Application to the GLC-2000 and MODIS land cover products. *IEEE Transactions on Geoscience and Remote Sensing*, *44*, 1740–1746.

Steele, B. M., Winne, J. C., & Redmond, R. L. (1998). Estimation and mapping of misclassification probabilities for thematic land cover maps. *Remote Sensing of Environment, 66*(2), 192–202.

Stehman, S. V. (2000). Practical implications of design-based sampling inference for thematic map accuracy assessment. *Remote Sensing of Environment, 72*, 35–45.

Stehman, S. V. (2006). Design, analysis, and inference for studies comparing thematic accuracy of classified remotely sensed data: a special case of map comparison. *Journal of Geographical Systems, 8*, 209–226.

Strahler, A. H., Boschetti, L., Foody, G. M., Friedl, M. A., Hansen, M. C., Herold, M., et al. (2006). Global land cover validation: Recommendations for evaluation and accuracy assessment of global land cover maps. *Technical Report, Joint Research Centre, Ispra, EUR 22156 EN* 51 pp.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography, 46*, 234–240.

Wang, F. (1990). Fuzzy supervised classification of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing, 28*, 194–201.

Wang, Q., Ni, J., & Tenhunen, J. (2005). Application of a geographically-weighted regression analysis to estimate net primary production of Chinese forest ecosystems. *Global Ecology and Biogeography, 14*, 379–393.

Woodcock, C. E., & Gopal, S. (2000). Fuzzy set theory and thematic maps: accuracy assessment and area estimation. *International Journal of Geographical Information Science, 14*, 153–172.