Geographically weighted accuracy for hard and soft land cover classifications: 5 approaches with coded illustrations

Alexis Comber & Naru Tsutsumida

Published online: 16 Oct 2023.

Submit your article to this journal ↗

Article views: 128

View related articles ↗

CrossMark View Crossmark data ↗

Geographically weighted accuracy for hard and soft land cover classifications: 5 approaches with coded illustrations

Alexis Comber [ID][a] and Naru Tsutsumida [ID][b]

[a]School of Geography, University of Leeds, Leeds, UK; [b]Department of Information and Computer Sciences, Saitama University, Saitama, Japan

ABSTRACT

This paper examines different geographically weighted (GW) approaches for calculating spatially distributed measures of accuracy / uncertainty, consolidating current approaches and proposing 2 new ones. GW frameworks use a moving window or kernel to extract and weight data subsets, from which local (ie spatially distributed) statistics or metrics are calculated. A validation dataset with hard and soft classifications is used to illustrate the approaches. It contains observed field survey data (also commonly derived from higher resolution imagery), and predicted data from a fuzzy c-means classification. The hard classes were used to estimate spatially distributed measures of overall, user's and producer's accuracies in two ways. First, by conceptualising them as probabilities to be estimated from generalised linear regression models (GLMs), extended into Geographically Weighted GLMs. Second, by constructing local GW correspondence matrices and then calculating local accuracy measures from these. The soft classes were used to calculate per-class measures of fuzzy certainty from the absolute difference between predicted and observed fuzzy memberships. Then, a novel fuzzy certainty logic is proposed and used to create fuzzy confusion matrices and per-class measures of fuzzy omission and commission error, supporting measures of fuzzy user's and producer's certainties. These were extended to the GW case to generate spatially distributed measures. Finally, the soft classifications were conceptualised as compositional data and measures of difference were estimated using Aitchison distances. In each case, the local hard and soft accuracy and certainty measures were interpolated over a 1 km grid to estimate accuracy surfaces. The context for this review is the increasing operational use of training and validation data, often with high numbers of records, containing both hard and soft classes. The data and R code used to undertake all the analyses in this paper are provided, supporting more nuanced analyses of such data.

CONTACT Alexis Comber ✉ a.comber@leeds.ac.uk 📧 School of Geography, University of Leeds, Leeds LS2 9JT, UK

1. Introduction

Without accuracy assessment the map is just a pretty picture, an untested hypothesis (verbatim quote of Giles Foody, International Symposium on Spatial Data Quality, La Grande Motte, France, 29 September 2015, but probably referencing McRoberts (2011))

This paper synthesizes geographically weighted (GW) methods for reporting and mapping the spatial distribution of classified remotely sensed (RS) data accuracy. It combines a number of inter-related concepts around accuracy in remote sensing and GW models:

- Classification accuracy in remote sensing provides a measure of the degree to which a sample of the classified data agrees with the reference data (Congalton 1991).
- Sample locations (pixels) should be selected under a probability-based sample design (Olofsson et al. 2014).
- At those locations, accuracy is quantified by comparing the *predicted* class (as generated by the classification algorithm applied to the RS data) (e.g. k-means, c-means, random forest, etc.) with the *observed* reference class that is deemed to be of higher quality in some way. The observed data are often from higher resolution imagery (Tsutsumida et al. 2020).
- Measures of overall accuracy (OA), producer's accuracy (PA) and user's accuracy (UA) are generated from cross tabulations of predicated against observed class. These describe probabilities and as such can also be generated from generalized linear regression models (GLMs) (Comber 2013).
- Standard accuracy measures provide only global or *whole map* (Openshaw 1996) accuracies and do not describe any spatial variation in accuracy. However, it is well known that errors are frequently concentrated in certain areas (i.e. are localized) (Congalton 1988).
- GW models generate spatially distributed outputs. The best known is geographically weighted regression (GWR – Brunsdon, Fotheringham, and Charlton (1996)). The GW framework uses a moving window to extract and weight data subsets, and create a series of local models, that are combined into a GW model (Gollini et al. 2015).
- GWR has a GLM form that can be used to generate geographically weighted generalized linear regression models (GW-GLM) (Nakaya 2015) and the GW framework can also be used to calculate other user defined local metrics (e.g. Comber and Harris (2018)).

This paper links these concepts and describes methods for generating local measures of spatially distributed accuracy and certainty to augment current approaches. It extends from hard or crisp classifications, in which each image pixel or object is allocated to a single class, to soft or fuzzy classifications. These contain pixel-level probabilities (or fuzzy class memberships) for each class, potentially indicating sub-pixel composition or classification uncertainty. The paper draws together developments in GW accuracy, describes two novel approaches for quantifying soft classification accuracy (fuzzy accuracy and compositional accuracy) and applies these under GW interpolation frameworks. The context for this work is the increasing interest in local accuracy measures – how and where classification accuracy varies – and the enhanced ability to generate such measures

due to increased use of high-resolution data to train and validate coarser resolution RS data of greater coverage. Such training/validation data overcomes concerns about the value of interpolated accuracy measures when estimated from a small number of observations (Foody 2022). Each approach is explored using the same validation dataset. The data and R code used to generate all the results and figures are at https://github.com/lexcomber/GWAcc.

2. Data

A single land cover validation dataset is used to demonstrate the approaches. It covers an area around Tripoli in Libya as described in Comber et al. (2012), although it has been subsequently modified. It contains the *predicted* hard land cover class and soft land cover class proportions from a fuzzy *c-means* classification (Wang 1990) of SPOT (Système Pour l'Observation de la Terre) resampled to 30 m for 210 locations. It also contains the *observed* hard class and soft class proportions at each location collected via field survey as the reference data. In order to generate spatially distributed measures of accuracy, a 1 km interpolation grid was defined to provide a spatial framework over which local accuracy measures were estimated. This size was arbitrarily chosen to provide a balance between spatial detail and computational efficiency. The validation data locations and the 1 km interpolation grid are shown in Figure 1.

2.1. Hard classification data

The cross-tabulation of the predicted and observed hard classes is shown in Table 1, with OA, a kappa estimate, and UAs and PAs for each of the five classes. This is variously referred to as the error matrix, the confusion matrix or the correspondence matrix. OA describes the proportion of observations correctly predicted. UA is the complement of the Commission error of the remote sensing class. It is the probability that the class predicted by the remote sensing analysis will be found in the observed data. PA is the complement of the Omission error of the remote sensing class. It describes the probability of the true
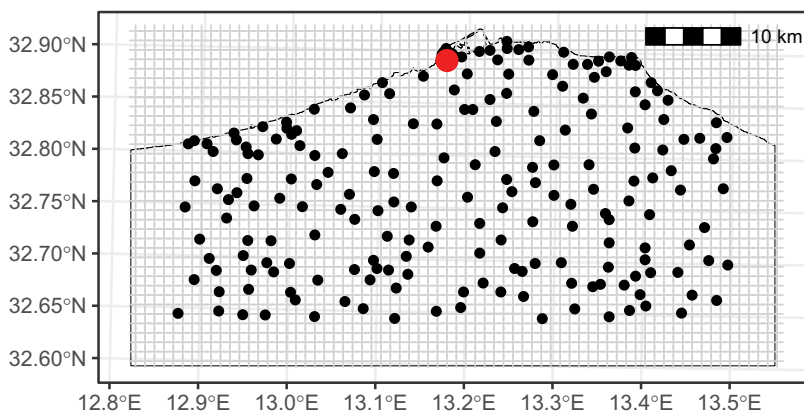

Figure 1. The locations of the validation points, a 1 km interpolation grid and Tripoli highlighted in red.

Table 1. The correspondence matrix between predicted (rows) and observed (columns) land cover class, with user's and producer's accuracies. Overall accuracy: 0.733; kappa estimate: 0.665.

	B	G	U	V	W	User's	Producer's
1. Bare Ground	19	5	5	4	3	0.528	0.864
2. Grazing Land	2	30	0	12	9	0.566	0.769
3. Urban	1	0	35	2	2	0.875	0.778
4. Vegetation	0	3	4	42	1	0.840	0.689
5. Woodland	0	1	1	1	28	0.903	0.651

observed class being found in the remote sensing analysis. Kappa estimates of accuracy provide an overall measure of accuracy informed by the probabilistic agreement and the expected disagreement between predicted and observed classes, although they are now somewhat discredited (Foody 2020; Pontius and Millones 2011). An overview of these measures derived from the correspondence matrix is comprehensively described in Congalton (1991).

2.2. Soft classification data

Soft classifications provide a measure of the uncertainty in the classification. A standard classification generates probabilities of the pixel belonging to each class based on the distance of the pixel to each class centre (whether unsupervised or supervised). Alternatively, soft classification generates measures of the partial membership of each pixel to each class, under a fuzzy approach (Fisher and Pathirana 1990), or sub-pixel proportions under a compositional approach (Wang, Shi, and Atkinson 2014). The validation dataset contains *observed* and *predicted* sets of soft classifications. These are shown in Figures 2 and 3 respectively. There are large differences in the spatial distributions of the predicted and observed soft class values caused by differences in how they were recorded: continuous predicted values were from a fuzzy *c-means* classification and the observed values were from a field survey in which class proportions were recorded in units of $\frac{1}{16}$ of a pixel.
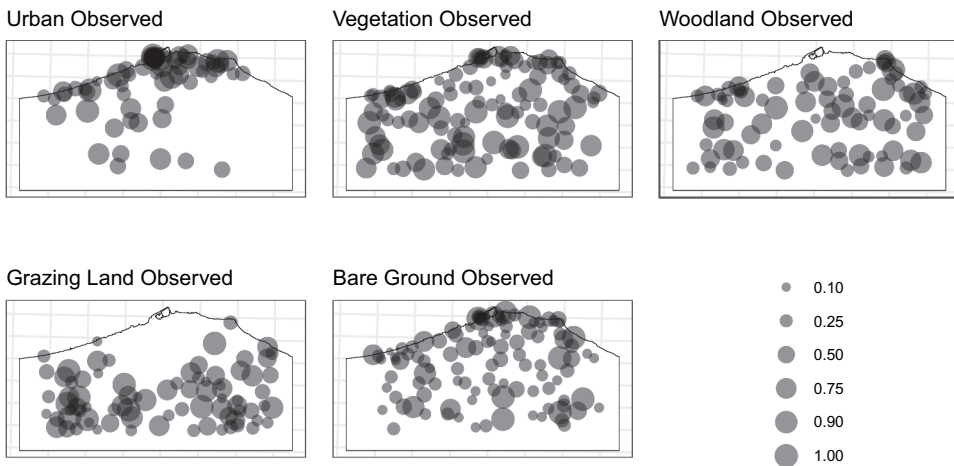

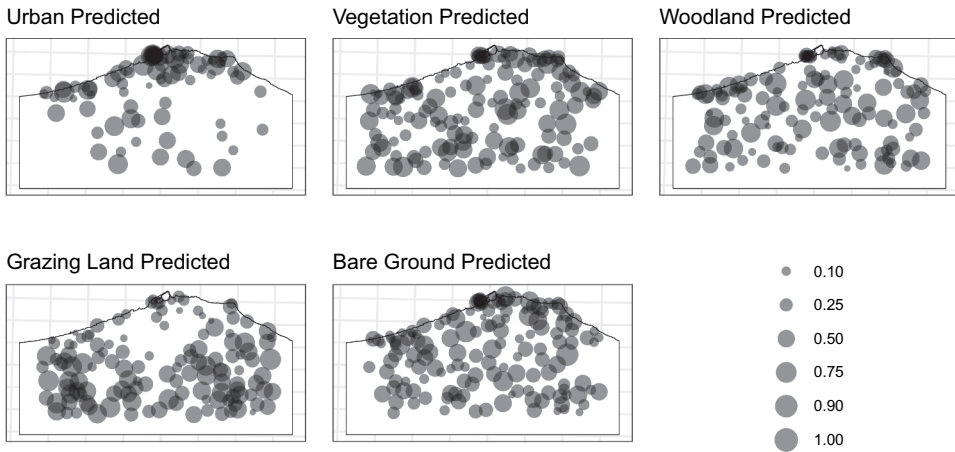**Figure 2.** The observed soft class memberships.

Figure 3. The predicted soft class memberships.

3. Working with hard classes

3.1. Accuracy as GLM and GW-GLM probabilities

Overall, user's and producer's accuracies are probabilities and can therefore be derived from generalized linear regression models (GLMs) with logistic form (Comber 2013). This is because of the allocation of each pixel to one observed class and one predicted class, allowing the validation data to be recast as True (1) or False (0) and thus probabilistic accuracy measures to be calculated.

OA can be estimated using a GLM, applied to a single vector of 1s indicating the same class is present in both the observed and validation data, 0s otherwise. This is regressed against an explanatory variable of 1 (i.e. an intercept term) and returns a single coefficient estimate, B_0. The OA probability is computed from antilogs (*exp*) of B_0 generated by the logistic GLM, as follows:

$$P(OA = 1) = \frac{exp(\beta_0)}{1 + exp(\beta_0)} \tag{1}$$

Effectively this is the probability that the observed class and the predicted class are the same. It can be derived from the cross tabulation in Table 1, where the 154 diagonal elements are indicated as 1 out of total count of 210, with the 56 off-diagonal elements indicated as 0. The result is a probability that observed class equals predicted class (i.e. OA) of 0.733 (154 / 210).

Equation 1 can be written in a different way by defining a *logit* function to transform any coefficient β (Equation 2), creating a regression model (Equation 3), and then applying the *logit* transform to calculate OA (Equation 4):

$$logit(\beta) = \frac{exp(\beta)}{1 + exp(\beta)} \tag{2}$$

$$ln(\frac{p}{1-p}) = \beta_0 \tag{3}$$

$$P(OA = 1) = logit(\beta_0) \tag{4}$$

where, p is the probability that $OA = 1$, and β_0 is the intercept term as before. The use of Equation 2 and the form of Equation 4 are important, as below they will be expanded to describe GW models.

For UA and PA, consider a subset of the validation data pertaining to the class of Grazing Land. It can be represented by two similarly ordered binary vectors of 210 elements, with the observed (reference) vector having 39 elements indicated as True (1) and the predicted (classified) vector having 53. User's accuracy can be estimated as follows:

$$P(y = 1) = logit(\beta_0 + \beta_1 x_1) \tag{5}$$

where $P(y = 1)$ is the probability that the observed class, y, is correctly predicted by the classified data, x_1, β_0 is the intercept term and β_1 is the slope. The result describes the probability that the observed data equals 1, given that the predicted data equals 1, which in this case is 0.57.

Inverting the response and explanatory variables in Equation 3 calculates producer's accuracy:

$$P(x = 1) = logit(\beta_0 + \beta_1 y_1) \tag{6}$$

where $P(x = 1)$ is the probability that the classified class is correctly predicted by the observed data, y_1, β_0 is the intercept term and β_1 is the slope. The result describes the probability that the predicted data equals 1, given that the observed data equals 1 and the producer's accuracy for Grazing Land is 0.77.

Having outlined the calculation of accuracy probabilities derived from the cross-tabulation using logistic regression models, it is now possible to consider the *geographically weighted* extensions to these. Geographically weighted regression (GWR) is similar in form to ordinary regression but uses a moving window or kernel under which local models, such as a logistic regression, are computed (Brunsdon, Stewart Fotheringham, and Charlton 1996; Fotheringham, Brunsdon, and Charlton 2002; Nakaya et al. 2005). This can be done at discrete locations through the study areas such as the validation locations as in Figure 1, but it is often more informative to estimate these over an interpolation surface such as the 1 km grid shown in Figure 1. The GWR form of the regression models described in Equations 4, 5, and 6 were extended to the spatial case in the following way:

$$P(OA = 1)_{(u_i,v_i)} = logit(\beta_{0(u_i,v_i)}) \tag{7}$$

$$P(y = 1)_{(u_i,v_i)} = logit(\beta_{0(u_i,v_i)} + \beta_1 x_{1(u_i,v_i)}) \tag{8}$$

$$P(x = 1)_{(u_i,v_i)} = logit(\beta_{0(u_i,v_i)} + \beta_1 y_{1(u_i,v_i)}) \tag{9}$$

where (u_i, v_i) are vectors of two-dimensional coordinates describing the location of each observation i in the validation dataset or the interpolation grid.

In a GW approach, a moving window or *kernel* is used to generate weighted data subsets from which local models are constructed at discrete locations: the data are subsetted and weighted such that data points further away from the kernel centre contribute less to the local model. 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 kernel to (u_i, v_i). A typical kernel function is the *bisquare* kernel – see Gollini et al. (2015) for details of kernels in GW models. For a given bandwidth b, this is defined in Equation 10.

$$w_{i,j} = \begin{cases} \left(1 - \left(\frac{d_{i,j}^2}{b^2}\right)\right) & \text{if } d_{ij} < b; \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

where $d_{i,j}$ is the distance from the kernel centre to the data point j and the weights $w_{i,j}$ for each data point i change for each location.

It is important to note that all GW models require a kernel bandwidth. This describes the size of the moving window (fixed distance bandwidth) or the proportion of the data points used in each local calculation (adaptive distance bandwidth). The bandwidth defines the degree of smoothing in the outputs. All implementations of GWR have procedures for bandwidth optimization to determine the bandwidth that optimizes error or fit, such as leave-one-out-cross validation or metrics like AIC (Akaike 1998). In this paper, bi-square kernels with adaptive distances (specifying a proportion of data points rather than fixed distances) were used throughout and these were optimized to minimize root mean square prediction error. Bandwidths are reported in proportions throughout the paper, and where percentages (of data points) are used, these are clearly indicated. Geographically weighted GLMs (GW-GLMs) were implemented using the spgwr R package (Bivand et al. 2020).

Summaries of the distributions of the local GW-GLM derived OA, UAs and PAs, and their adaptive bandwidths (expressed as proportion of the total number of data points), are shown in Tables 2, 3 , and 4. The global accuracy measures from Table 1 are included for

Table 2. A summary of the local measures of overall accuracy generated by the GW-GLM models, with the adaptive bandwidth and the global overall accuracy measure.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Bandwidth	Global
0.649	0.705	0.746	0.74	0.777	0.814	0.819	0.733

Table 3. A summary of the local measures of user's accuracy generated by the GW-GLM models, with the adaptive bandwidths and the global user's accuracy measures.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Bandwidths	Global
1. Bare Ground	0.489	0.502	0.506	0.519	0.542	0.559	1.000	0.528
2. Grazing Land	0.453	0.518	0.597	0.611	0.671	0.912	0.624	0.566
3. Urban	0.641	0.842	0.887	0.861	0.909	0.929	0.681	0.875
4. Vegetation	0.770	0.823	0.840	0.839	0.857	0.880	0.995	0.840
5. Woodland	0.861	0.869	0.891	0.901	0.937	0.952	1.000	0.903

Table 4. A summary of the local measures of producer's accuracy generated by the GW-GLM models, with the adaptive bandwidths and the global producer's measures.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Bandwidths	Global
1. Bare Ground	0.855	0.862	0.866	0.866	0.869	0.883	1.000	0.864
2. Grazing Land	0.117	0.648	0.760	0.748	0.876	0.978	0.376	0.769
3. Urban	0.592	0.730	0.763	0.755	0.797	0.826	0.833	0.778
4. Vegetation	0.620	0.686	0.705	0.708	0.727	0.812	0.843	0.689
5. Woodland	0.504	0.624	0.658	0.669	0.719	0.818	0.800	0.651

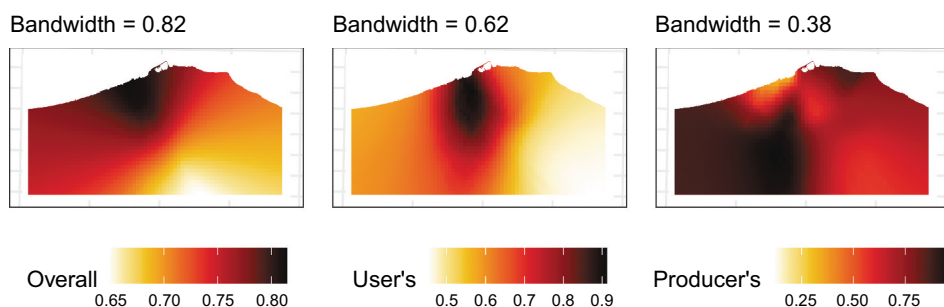Bandwidth = 0.82   Bandwidth = 0.62   Bandwidth = 0.38

Figure 4. Spatially distributed overall accuracy, and user's and producer's accuracies for Grazing Land from a GW-GLM.

comparison. There are a number of interesting observations when the bandwidths and the distribution summaries are examined:

- Overall accuracy (Table 2) has some degree of spatial variation and high bandwidth indicating that 82% of the observations were used in each local OA model. The inter-quartile range indicates the spatial homogeneity of OA in this case.
- For user's accuracy (Table 3), there is very little variation in accuracy for Bare Ground, Vegetation Woodland as reflected in the narrow inter-quartile ranges and large bandwidths all around 100% of the data points. (Note that in these cases the spatial variation is driven by the weighting function). The UAs for Grazing Land show some spatial variation with important differences in its inter-quartile ranges, suggesting that this accuracy are highly localized. Urban has a narrow range and a small bandwidth, suggesting some spatial variation.
- The producer's accuracies (Table 4) are interesting. Here, the bandwidth does not universally indicate the amount of spatial heterogeneity in the accuracy measure: despite moderately high bandwidths (80% of the data points) there is considerable variation in Urban and Woodland as well as in Grazing Land (low bandwidth).

The spatial variation of the GW accuracy measures can also be further explored by mapping them. OA with UA and PA for Grazing Land are shown in Figure 4. Here, higher values of PA and UA are found in the centre and north of the study area, areas with low levels of observed and predicted Grazing Land, and areas of higher PA are found in the south and west.

3.2. GW correspondence matrices

An alternative approach for estimating local measures of OA, UA and PA is to construct *local* correspondence matrices (cross-tabulations) of observed and predicted values. Comber et al. (2017) describe the use of geographically weighted correspondence matrices (GWCM), with functionality provided by the gwxtab R package (https://github.com/chrisbrunsdon/gwxtab). Here, a GWCM was constructed at each location in the 1 km grid from which probability-based accuracy measures were then computed directly.
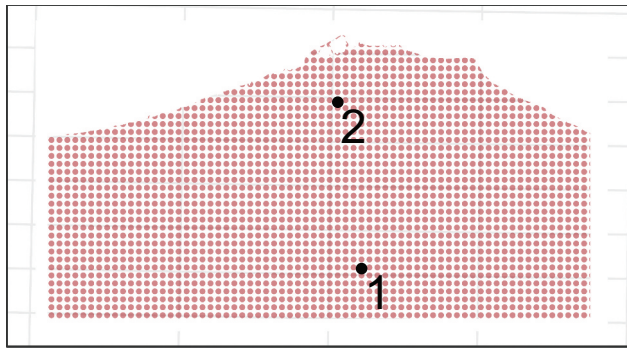
Figure 5. Two example locations for which geographically weighted grids are determined.

Table 5. The GWCM for point 1, with predicted (rows) and observed fuzzy class (columns). The values indicate GW pixel counts.

	B	G	U	V	W
1. Bare Ground	1.83	1.16	0.03	0.00	0.63
2. Grazing Land	0.00	3.61	0.00	1.41	1.10
3. Urban	0.00	0.00	0.60	0.04	0.96
4. Vegetation	0.00	0.28	0.93	4.84	0.26
5. Woodland	0.00	0.00	0.00	0.89	1.04

The GWCM approach, through the gwxtab package, is to construct spatial cross tabulation *functions* that are parameterized with the predicted and observed data, and then used to create local GWCMs for a given bandwidth and location. This process is best illustrated through examples. Figure 5 shows two locations on the 1 km of the grid. The spatial cross-tabulation function generates geographically weighted correspondence matrices for those locations as shown in Tables 5 and 6 using an adaptive bandwidth of 30% of the data points. These show the geographically weighted cross-tabulation counts of pixels at each location. Note that the bandwidth for GWCM construction cannot be optimized and instead has to be selected by the user.

The bandwidths determined in the GW-GLM approach above and the interpolation grid were passed to the GWCM function to generate GW correspondence matrices at each location from which local measures of GW OA, UA, PA and kappa accuracy were calculated. Summaries of the distributions of local geographically weighted OA and kappa accuracy are shown in Table 7, and the GW UAs and PAs for each class are in Tables 8 and 9 respectively. The OA, UA and PA results are exactly the same as those from the GW-GLM in Tables (2–4). Although the kappa estimate has been criticized for its formulation, its embedded

Table 6. The GWCM for point 2, with predicted (rows) and observed fuzzy class (columns). The values indicate GW pixel counts.

	B	G	U	V	W
1. Bare Ground	2.15	0.00	1.68	0.05	0.00
2. Grazing Land	0.00	0.15	0.00	0.00	0.00
3. Urban	0.64	0.00	10.94	0.55	0.91
4. Vegetation	0.00	0.14	0.44	4.41	0.00
5. Woodland	0.00	0.00	0.59	0.00	3.63

Table 7. The local overall and kappa accuracies from a GWCM under an adaptive bandwidth of 0.82, with the global statistic.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Global
Overall	0.649	0.705	0.746	0.740	0.777	0.814	0.733
Kappa	0.543	0.627	0.675	0.669	0.717	0.757	0.665

Table 8. A summary of the local measures of user's accuracy generated by the GW correspondence matrix approach, with the adaptive bandwidths and the global user's accuracy measures.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Bandwidths	Global
1. Bare Ground	0.488	0.501	0.504	0.519	0.543	0.560	1.000	0.528
2. Grazing Land	0.453	0.518	0.597	0.611	0.671	0.912	0.624	0.566
3. Urban	0.641	0.842	0.887	0.861	0.909	0.929	0.681	0.875
4. Vegetation	0.770	0.823	0.840	0.839	0.857	0.880	0.995	0.840
5. Woodland	0.860	0.866	0.889	0.900	0.937	0.952	1.000	0.903

Table 9. A summary of the local measures of producer's accuracy generated by the GW correspondence matrix approach, with the adaptive bandwidths and the global producers's accuracy measures.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Bandwidths	Global
1. Bare Ground	0.855	0.862	0.866	0.866	0.869	0.883	1.000	0.864
2. Grazing Land	0.117	0.648	0.760	0.748	0.876	0.978	0.376	0.769
3. Urban	0.592	0.730	0.763	0.755	0.797	0.826	0.833	0.778
4. Vegetation	0.620	0.686	0.705	0.708	0.727	0.812	0.843	0.689
5. Woodland	0.504	0.624	0.658	0.669	0.719	0.818	0.800	0.651

assumptions and difficulties in its interpretation (Foody 2020; Pontius and Millones 2011), it is included here to illustrate how non-GLM related metrics can also be calculated from the correspondence matrix.

3.3. Summary of GW accuracy for hard classes

Two approaches for generating GW measures accuracy are described, based on local GW GLMs and local GW correspondence matrices. Both generate the same results – compare Tables 3 with 8 and Tables 4 with 9. The logic for this equality is as follows: the weighted subsets of binary variables that are passed to the GW-GLM are effectively the inputs into a 2 by 2 correspondence table of counts for each class. The locally weighted subsets of these used to create each local model in the GW-GLM approach are effectively the same subsets that are used to construct the local GW correspondence matrices, given the same bandwidth and weighting function.

Each approach has a distinct advantage over the other. The GWCM is more flexible as it can accommodate any measure derived from the correspondence matrix. For example, Pontius and Millones (2011) and Pontius Jr and Santacruz (2014) describe a number of metrics including composite measures of accuracy. These have been included in the differR package (Pontius and Santacruz 2015) and applied in the paper by Comber et al. (2017). However, unlike GW-GLM approach, there is not an obvious method to determine optimal bandwidths for the GWCMs. This is a critical point. It means that the scales of heterogeneity have to be determined elsewhere (such as in a GW-GLM) or are user defined, which is problematic. By way of example, Figure 6 shows two spatially varying
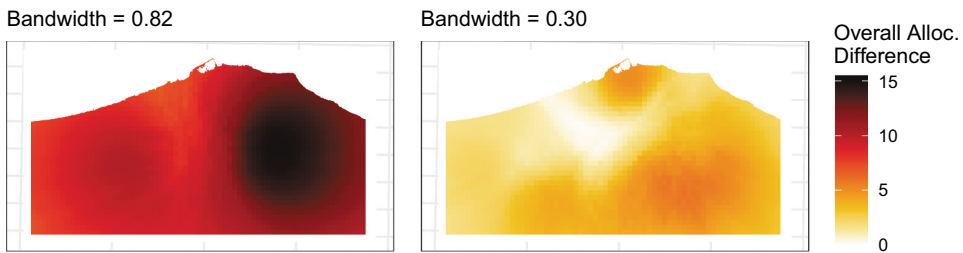
Figure 6. Two examples of overall allocation difference estimates from GW correspondence matrices with different bandwidths.

overall allocation difference measures (Pontius and Santacruz 2015) generated under an adaptive kernel using 30% of the data points (a reasonable configuration in many GW analyses) and an adaptive bandwidth using 82%, determined optimally for the GW-GLM OA measure. These two maps provide very different representations of the spatial variation of the same accuracy measure, with the larger bandwidth generating a higher values and potentially a spatially smoother results. One obvious option is to determine bandwidths using a GW-GLM and then use the GWCM to calculate an appropriate bandwidth for the desired local metric (if it is not OA, UA or PA).

4. Working with soft classes

In hard classifications, each pixel is uniquely defined as belonging to one class only. There are many embedded assumptions associated with hard classification and Fisher (1997) provides a thoughtful discussion of these. Soft classifications contain measures of proportion, membership or probability for each class, summing to 1 for each pixel. They capture some of the uncertainty associated with hard classifications, which may be due to the mismatch between pixel resolution and the granularity of the process being monitored (i.e. sub-pixel processes) or due to the differences in the spectral characteristics of the pixel relative to the classes. By capturing the uncertainty and the inherent properties of the image pixel in remote sensing outputs, soft classifications can provide more representative measures of the real world. As a result, the remote sensing community is increasingly interested in working with the uncertainty captured by soft classifications (Alshari and Gawali 2021; Atkinson 2005; Khatami, Mountrakis, and Stehman 2017b). In the subsections below, the pixel is considered as a fuzzy object, with partial membership to each class and then as compositional data, indicating the fractions of different classes within the pixel. The concept of compositional data is well developed in some areas of science and is starting to gain traction in the remote sensing community (Zakeri and Mariethoz 2021).

4.1. GW fuzzy certainty I

A fuzzy classification is one that allows for classified objects (pixels or parcels) to have partial memberships to the set of all possible classes. These are similar to the probabilities generated, for example, from a *k-means* or maximum likelihood classifier, and are often

Table 10. Summaries of the distribution of the GW fuzzy certainty for each class, with optimized kernel bandwidths.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Bandwidths
1. Bare Ground	0.722	0.887	0.929	0.913	0.953	0.978	0.076
2. Grazing Land	0.744	0.790	0.823	0.832	0.868	0.941	0.076
3. Urban	0.722	0.887	0.929	0.913	0.953	0.978	0.076
4. Vegetation	0.818	0.824	0.833	0.834	0.844	0.851	0.405
5. Woodland	0.865	0.865	0.869	0.868	0.871	0.872	0.776
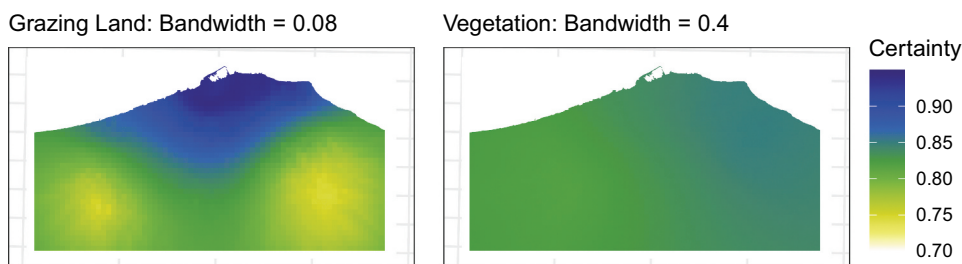
Figure 7. The spatial distributions of the fuzzy certainty of Grazing Land and Vegetation.

used as proxies for fuzzy class memberships. The critical advance by Comber et al. (2012) was to create Geographically Weighted measures of fuzzy accuracy or more correctly fuzzy certainty. They used a GW framework to compute local measures of difference for each class at each location. If y is the observed fuzzy membership and x is the predicted fuzzy membership, at any given location (u_i, v_i), the fuzzy certainty for any given class $\mu(z)$ can be obtained from the complement of the absolute difference between the two fuzzy memberships:

$$\mu(z)_{(u_i,v_i)} = 1 - abs(y_{(u_i,v_i)} - x_{(u_i,v_i)}) \tag{11}$$

The quantity $\mu(z)$ can take any value in the range 0 to 1, with 0 corresponding to a hard classification agreement and 1 corresponding to hard classification disagreement between the observed reference classes and the predicted mapped classes.

The values of $\mu(z)$ for each location are regressed against unity (1) using an OLS regression under a GW kernel over the 1 km interpolation grid. This is a standard GWR (Brunsdon, Stewart Fotheringham, and Charlton 1996), generating coefficient estimates that describe the pre-class spatial variation in fuzzy certainty. The kernel bandwidth can be optimized and the results are summarized in Table 10. What is interesting is the high degree of correspondence between the predicted and observed fuzzy measures and the relatively narrow spatial variation of these irrespective of the bandwidth, which in some cases are very small (8% of the data). It is possible to map these as is done for the Grazing Land and Vegetation classes in Figure 7, which have very different optimized bandwidths and distributions. When these are considered in the context of the distributions in Figures 2 and 3, Grazing Land has low fuzzy certainty values in areas of dense predicted and observed validation data, whereas this is not the case for Vegetation, which has much lower spatial variation in fuzzy certainty values.

4.2. GW fuzzy certainty II

It is possible to take a more informed and nuanced approach to fuzzy accuracy by extending the logic proposed by Fisher et al. (2006) for *fuzzy change* into *fuzzy accuracy*. The change matrix is a direct corollary of the correspondence matrix as it compares two sets of observations collected at the same locations. Fisher et al. (2006) describe the construction of a fuzzy change matrix from the class memberships values observed at Time 1 and Time 2 and evaluate fuzzy change using the concept of bounded difference to determine per-class measures of fuzzy loss and fuzzy gain. If the change matrix is composed of rows describing classes at Time 1 and columns at Time 2 then fuzzy loss equates to the sum of the row-wise off diagonal elements for each class, and fuzzy gain to the sum of column-wise off diagonal elements. These concepts of change from Fisher et al. (2006) can be extended to fuzzy accuracy and uncertainty as follows:

- Fuzzy producer's **certainty** is the complement of fuzzy Omission error, defined as fuzzy memberships that were not recorded (omitted) in the predicted fuzzy memberships from the remote sensing analysis. It is calculated from 1 - (fuzzy Omission error) from the column-wise off diagonal elements in the fuzzy correspondence matrix, for each class. This is the corollary of fuzzy gain.
- Fuzzy user's **certainty** is the complement of fuzzy Commission error, defined as fuzzy memberships recorded in the predicted fuzzy memberships but not found in the observed memberships from the field survey. It is calculated from 1 - (fuzzy Commission error) from the row-wise off diagonal elements in the fuzzy correspondence matrix for each class. This is the corollary of fuzzy loss.

These fuzzy certainties can be determined by adapting the logics of bounded difference, fuzzy loss and fuzzy gain. It also is possible to extend these to a GW framework to generate spatially distributed measures. This is entirely novel and is reported for the first time in this paper.

The fuzzy change analysis in Fisher et al. (2006) reasons with continuously valued fuzzy maps that represent each land cover class at T_1 (Time 1) and at T_2 (Time 2). These are recorded as a real number, μ, between 0 and 1, indicating the degree to which each object (pixel) matches a land cover type in each time period. Fuzzy change can be recast as fuzzy certainty, such that the mapped *predicted* membership from the remote sensing classification are compared with the reference *observed* land cover fuzzy memberships from the field survey. Following Fisher et al. (2006) we can denote these as *Pred* and *Obs* (predicted and observed memberships), instead of T_1 and T_2, respectively. To determine certainty measures, a novel fuzzy set logic has to be established. For the fuzzy certainty matrix, the logic of populating the cells in the fuzzy correspondence matrix is variable.

For the leading diagonal elements, the logic can be expressed as those elements where the land cover class is C_i in *Pred* and C_i at *Obs*: i.e. the intersection of the two fuzzy sets. The normal intersection operator for fuzzy sets takes the minimum of the two fuzzy membership values, μ (Zadeh 1965), as in Equation 12.

$$\mu(C_iPred, C_iObs) = min(\mu(C_iPred), \mu(C_iObs)) \tag{12}$$

For example, if the membership of C_i in *Pred* is 0.5 and C_i at *Obs* of 0.6, the membership of the intersect will be 0.5. As P. Fisher et al. (2006) notes with respect to fuzzy change, the use of the minimum operator '*is reasonable as the lower value records the amount unchanged*' (Fisher et al. 2006, 165).

The off-diagonal elements of the fuzzy correspondence matrix need to be treated in a different way and the logic is different. It involves consideration of two different fuzzy classes C_i in *Pred* and C_j in *Obs*. These elements can be interpreted as the intersection of the two fuzzy land cover types, C_iPred and C_jObs, where $i \neq j$, or as the intersection of the fuzzy producer's certainty in C_i and the fuzzy user's certainty in C_j across the two measurements, *Pred* and *Obs*. Fisher et al. (2006) undertake some investigation of different fuzzy loss and gain approaches in the context of fuzzy change and noted that the Bounded Difference between two fuzzy sets is '*the simplest operator for which the results make the most sense*' (p166) when used to compute fuzzy loss and fuzzy gain.

This approach to fuzzy change is transposed to fuzzy accuracy. Here, evaluation of the difference between observed and predicted fuzzy objects incorporates a boundary term, defined as the $min(\mu(C_1, Pred), \mu(C_2, Obs))$ for loss/Commission and $min(\mu(C_1, Obs), \mu(C_2, Pred)$ for gain/Omission. Fuzzy Commission Error (the complement of fuzzy user's certainty) and Fuzzy Omission Error (the complement of fuzzy producer's certainty) are defined with bounded difference as follows:

$$\mu(OmissionC_1) = max(0, \mu(\neg C_1, Pred) + min(\mu(C_1, Obs), \mu(C_2, Pred)) - 1) \quad (13)$$

$$\mu(CommissionC_1) = max(0, min(\mu(C_1, Pred), \mu(C_2, Obs)) + \mu(\neg C_1, Obs) - 1) \quad (14)$$

From these fuzzy producer's certainty and fuzzy user's certainty are taken from the complements of fuzzy omission and commission, respectively. Set operations defined in this way can be used to create the fuzzy correspondence matrix, with measures of per class fuzzy user's or Commission errors and producer's or Omission errors. If the predicted class membership *Pred* are rows in the correspondence matrix, then the upper off-diagonal elements in a fuzzy correspondence matrix describe fuzzy commission errors (false positives). Similarly, if the observed class membership *Obs* are columns in the correspondence matrix, then the lower off-diagonal elements in a fuzzy correspondence matrix describe fuzzy omission errors (false negatives).

The fuzzy correspondence matrix in Table 11 indicates the uncertainty in the soft validation data. This shows for example, that 6.2 pixels of Grazing Land in the predicted data from remote sensing (rows) were observed in the Bare Ground class (via field survey), and that 5.2 pixels observed in the field survey of Urban were found in the Woodland remote sensing class. The overall patterns are similar to those in Table 1, the standard

Table 11. Summaries of the distribution of the GW fuzzy certainty for each class, with optimized kernel bandwidths.

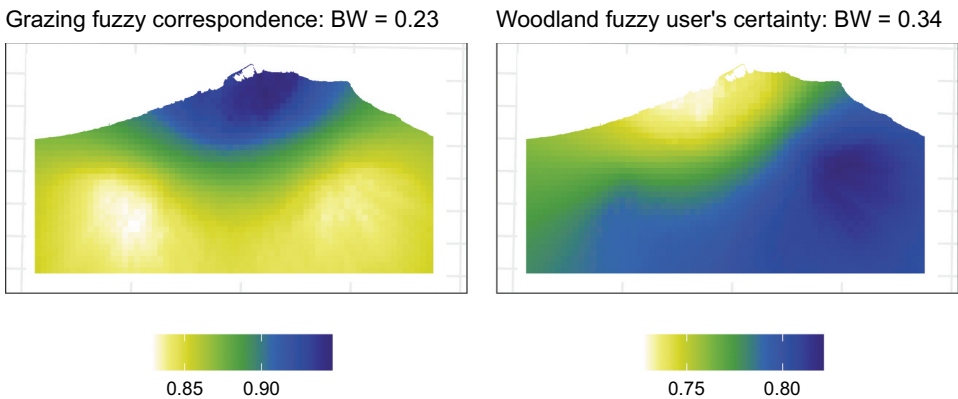|	B	G	U	V	W	Omission	Commission
1. Bare Ground	26.907	0.545	1.865	1.100	1.421	20.370	4.931
2. Grazing Land	6.220	26.840	0.250	0.918	1.197	11.723	8.585
3. Urban	1.714	1.154	22.198	2.744	0.803	10.775	6.415
4. Vegetation	5.377	4.651	3.449	35.275	2.040	10.375	15.516
5. Woodland	7.060	5.374	5.211	5.613	25.774	5.461	23.257

Figure 8. Examples of fuzzy correspondence and fuzzy user's certainties.

correspondence matrix, but with the uncertainty in the hard classification now explicitly represented. Note that the sum of values in the fuzzy correspondence matrix is 196, less than the 210 observations, due to the bounded difference operations as discussed in Fisher et al. (2006). The fuzzy omission and commission uncertainties summarized in Table 11 show the sums of the off-diagonal row and column elements of the fuzzy correspondence matrix.

It is possible to extend the measures from the fuzzy correspondence matrix to the GW case by extracting the per class fuzzy overall certainty, user's and producers' certainties from the measures of fuzzy correspondence, omission and commission, respectively. A GWR can be used to generate spatially distributed estimates of these over the 1 km interpolation grid. Investigation of the interpolated results indicate higher variations in fuzzy correspondence for Grazing Land and in fuzzy user's certainty for Woodland. These are shown Figure 8 with similarities to the fuzzy certainty in Figure 7 for Grazing Land but with a narrower range in the fuzzy correspondence. Generally, much lower variation was found than in the previous fuzzy certainty analysis, and little variation was found in the fuzzy producer's certainties.

4.3. GW compositional error

A final approach is to consider the soft classification values as compositional data. These are counts or measurements that are a proportion of some whole (Aitchison 2005). Examples include soil composition described in percentages of Sand, Silt, and Clay, measures of chemical concentrations in water samples in parts per million (ppm), or the proportions of the population in different age categories. In a remote sensing context, the class probabilities or memberships generated as part of a standard remote sensing image classification can be considered as compositional data. The soft class values are commonly conceived as the probability of each pixel belonging to each class, or in a fuzzy approach as the measure of the partial membership of the pixel to each class. In this case, the Observed and Predicted soft classes in the validation dataset can consider as compositional data because they are direct measures of the composition of the pixel area: the areal class membership from a field survey and from a fuzzy *c-means* classification

algorithm (Wang 1990). However, class probabilities can be treated in the same way (Tsutsumida et al. 2020).

There is a long literature on the analysis of compositional data. Much of it considers the difficulties in applying standard statistical approaches to such data, which frequently use Euclidean distances in a multi-variate feature space to generate similarity measures. These are confounded when calculated over a compositional attribute space, with issues associated with scale, perturbation, permutation variance and sub-compositional dominance. To overcome this, a number of different distance metrics have been proposed, one of which is Aitchison distance (Aitchison 1982). This was developed to support the measurement of difference between two compositions, such as the predicted and observed soft classes in the validation dataset. It has recently been applied to examine class proportions within the remote sensing domain (Tsutsumida et al. 2020; Zakeri and Mariethoz 2021) and is extended here. In overview, Aitchison distance is the Euclidean distance between *centred log-ratio* (*clr*) transformed compositions. When Euclidean distances are calculated from *clr*-transformed data, then the problems associated with standard statistical approaches are overcome. The *clr* transformation is undertaken as follows for a set of compositional data, *x*:

$$clr(x) = \left[log \frac{x_1}{g(x)} \dots log \frac{x_n}{g(x)} \right] \tag{15}$$

where x_1 to x_n are the counts or proportions in the compositional data and $g(x)$ is the geometric mean across all observations. If the observed and predicted class membership values are converted in this way, then the Aitchison distance in the two composite measures can be calculated using a standard Euclidean distance:

$$d_{Pred,Obs} = \sqrt{(clr(Obs) - clr(Pred))^2} \tag{16}$$

Aitchison distance requires non-zero values. To handle this, a small term is added to all observations in the data, in this case 0.01. In standard computational approaches, missing values are imputed in Aitchison distance analyses, but this is not correct for this data: zero data values are not missing, rather the feature was not found to be present. The addition of this term will change the value of the Aitchison distance and the results of the *clr* log transformation, but the key consideration is that compositional data reflect *relative* magnitudes and thus interest is in relative and not absolute differences. The robCompositions R package (Templ, Hron, and Filzmoser 2011) was designed specifically for the statistical analysis of compositional data. It includes an Aitchison distance function which undertakes the operations in Equations 15 and 16.

The Aitchison distance can be calculated for all locations collectively ($d_{Pred,Obs}$) and individually for each location i (d_{Pred_i,Obs_i}) which, if added up, sum to the overall Aitchison distance. These are shown in Table 12.

Table 12. A summary of the Aitchison distance distribution when calculated for each observation in the validation data, and the overall Aitchison distance value.

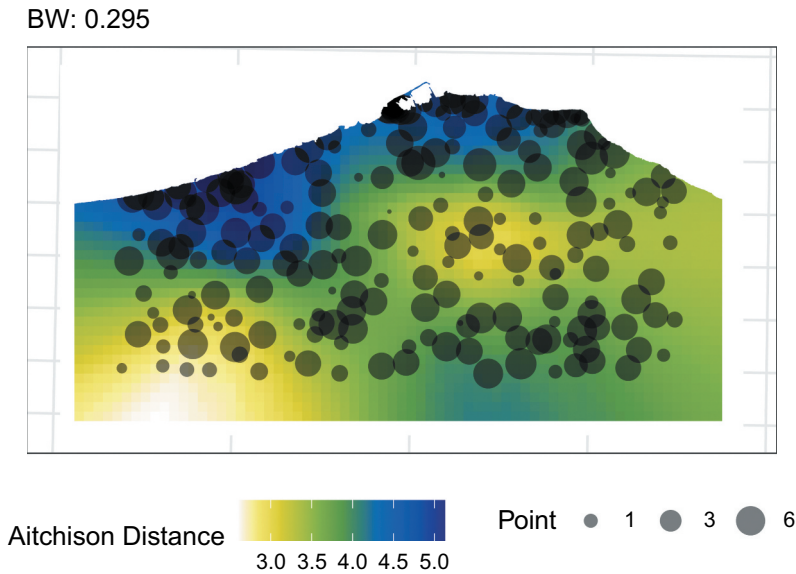| Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Overall
0.01	1.28	4.61	3.87	5.8	8.51	811.89

Figure 9. The Geographically weighted Aitchison distance (I), with the point values.

It is possible to generate a GW estimation of Aitchison distance locally by interpolating these measures over the 1 km grid using GWR. A GW kernel bandwidth was optimally determined and the result is mapped in Figure 9. This shows the lower distances between observed and predicted soft classifications to be in the south and west of the study area and to the centre immediately below Tripoli. The points indicate how the underlying data (Aitchison distance values at each of the 210 sample locations) generate these local variations in the interpolated surface. There are some similarities with the kappa distributions in Figure 6 under a bandwidth of 30% of the data points.

Figure 9 shows the interpolated Aitchison distance measure over the 1 km grid. An alternative is to calculate the Aitchison distance metric itself using a geographically weighted framework, as first done by Tsutsumida et al. (2020). In this local GW, Aitchison distance measures are computed from weighted data subsets at each location. Bespoke GW code was created to do this and the result is shown in Figure 10. This is somewhat different to Figure 9 with the differences in spatial pattern due to differences in the calculation of local Aitchison: Figure 9 shows interpolated values calculated from single data points, while Figure 10 shows values calculated from local data subsets. The difference in absolute values reflects the number of observations used to calculate each local Aitchison distance measure.

4.4. Summary of GW certainty for soft classes

The different approaches to soft classification have at their core an acknowledgement of the concept of the mixed pixel (Fisher 1997). They recognize that the area covered by a pixel may be composed of a mix of land covers or may have different properties to the exemplar classes in the training data. They do not accept the pixel allocation to the class with the largest sub-pixel area or largest probability. Instead, soft and compositional
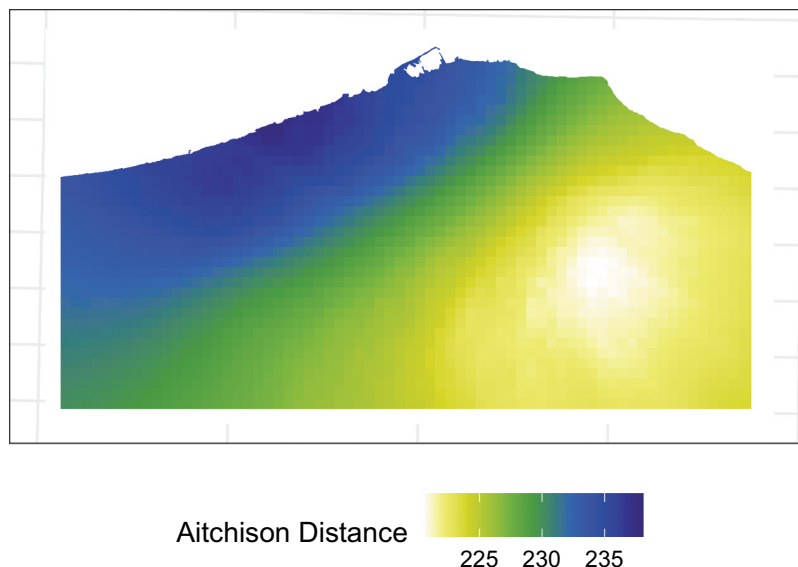
BW: 0.295

Figure 10. The Geographically weighted Aitchison distance (II).

approaches explicitly try to capture and reason with the classes that are present or probable, either through the idea of a pixel having partial probabilities for multiple classes or through the pixel being composed of multiple classes.

In the first fuzzy analysis, as reported in Comber et al. (2012), per-class surfaces of fuzzy certainties were generated. These were calculated from the absolute difference in fuzzy membership between predicted and observed for each class, for each validation location. The difference measures were then interpolated (regressed against 1) over the 1 km grid, using Geographically Weighted Regression (Brunsdon, Stewart Fotheringham, and Charlton 1996), for which an optimal bandwidth was determined (see Table 10), and the complement of the GW coefficient estimate was mapped (see Figure 7).

The second fuzzy analysis is entirely novel. It extended the application of fuzzy set theory to consider the difference between two fuzzy sets in relation to accuracy (or more correctly certainty), using the bounded difference and the fuzzy boundary. It transposed the logic of fuzzy change from Fisher et al. (2006) for quantifying fuzzy loss and fuzzy gain over time, to fuzzy certainty in order to quantify fuzzy commission and omission errors. These were used to construct a fuzzy correspondence matrix (Table 11) that summarized the classification uncertainty, and to map fuzzy correspondence and user's certainty shown for Grazing land and Woodland, respectively, in Figure 8.

A final and novel soft accuracy analysis considered the predicted and observed soft classes as fractions in a compositional analysis. The concept of compositional data is widely applied in other areas of science but has rarely been adopted in remote sensing, with only a few examples (e.g. Tsutsumida et al. (2020); Zakeri and Mariethoz (2021)). In compositional analysis, the aim to compare the overall composition of two observations or samples, and Aitchison distance (Aitchison 1982) provides a method for overcoming the problems with standard distance measures. This was calculated for each observation

and interpolated using a GWR as shown in Figure 9, showing very similar patterns to the fuzzy correspondence in Figure 8.

The critical decision about whether to adopt a fuzzy analysis over one grounded in compositional analysis, probably depends on whether the single fuzzy set (for example of Grazing Land) is the focus of the analysis or whether it is the overall composition of the observations that is of the primary interest (see Tsutsumida et al. (2019)).

5. Discussion

A number of approaches for generating spatially distributed accuracy from interpolated model-derived accuracies can be found in the literature. For example, Khatami, Mountrakis, and Stehman (2017a) investigated accuracy models of spatial and spectral properties and different interpolation functions, Ebrahimy et al. (2021) interpolated the outputs of a random forest model to predict mis-classifications from the spectral properties of the pixel relative to the class label, and in earlier work, Brown, Foody, and Atkinson (2009) investigated the effectiveness of different classification models in predicting accuracy and uncertainty. The effectiveness of such model-based approaches is driven by universal validation data characteristics: higher sample numbers generate better results and different classification algorithms and interpolation functions are more or less suited to the task in hand and the local environment. The same rubrics apply to the GW interpolation models described here.

This document has illustrated five different approaches for generating local measures of accuracy and certainty under some kind of geographically weighted framework. The approaches were grouped into to those that handle hard, single class allocations and those that accommodate soft, fuzzy, probable or compositional class measures. All of the approaches construct local measures of accuracy of some kind that were interpolated over a 1 km grid. The need for an interpolation framework is because of the desire to examine the spatial properties of accuracy over the entire study area and the nature of validation data sample. The validation data used in this paper are sparse, containing observations collected at only 210 locations but it provides a transparent walk-through of the different approaches. Many remote sensing accuracy assessments have a greater number of validation points containing the observed or true class as generated from higher resolution RS data (e.g. Tsutsumida et al. (2020)), and allowing soft land cover probabilities, memberships or compositions to be included in the reference data. This in turn provides opportunities to quantify the accuracy and uncertainty using the hard and soft GW approaches described in this paper, overcoming concerns about the value of interpolated accuracy measures when estimated from validation data with a small number of observations (Foody 2022).

Approaches for working with hard classifications (Section 3) are grounded in a conceptualization that sees each pixel as belonging to a single class, and the aim of classification is to allocate the pixel to the correct class. The objective in these cases is to determine the accuracy of that allocation. A GW framework was used to generate geographically weighted overall, user's and producer's accuracies, extending from the probabilities in the correspondence matrix, to a binomial GLM regression and then to a Geographically Weighted GLM regression (Comber 2013), generating local accuracy measures at each of the 210 validation data locations. The optimized GW bandwidths and

the interpolation of the local measures over a 1 km grid to provide explicit measures of the degree of spatial non-stationarity in the accuracy measures. The advantages of this approach is that it has clear links to a well-established theoretical framework for GLMs and to a degree for GWR-based approaches (Fotheringham, Brunsdon, and Charlton 2002). The limitation is that it can only be used to generate probability-related measures under a GLM and not other measures such as kappa.

The second part of Section 3 created local GW correspondence matrices at each validation location (Comber et al. 2017). However, there is no way to determine the optimal GW kernel bandwidth for the GWCM approach. This is a critical issue as the kernel bandwidth determines number of observations used in local model and thus the degree of smoothing in the outputs, with smaller bandwidths typically associated with greater local variation than larger ones. Here, the optimized bandwidths from the GW-GLM were passed to the GWCM functions and local measures of overall, user's, producer's and kappa accuracies were estimated. The overall, user's and producer's accuracies were shown to be the same as those generated using a GW-GLM. The advantages of GW correspondence matrices is that they allow the local form of any measure derived from the correspondence matrix to be calculated. For example, Comber et al. (2017) estimated a range of GW accuracy measures from Pontius Jr and Santacruz (2014) and Pontius Jr and Millones (2011). Its limitations are that bandwidths have to be specified and these need to be sufficiently large to avoid NA's values in the local correspondence matrices (for example, an adaptive bandwidth of 15% of the data points was found to fail for some local accuracy measures using this data).

Section 4 describes approaches for examining soft classification validation data. This represents a different conceptualization to that of hard classification. Hard approaches inherently assume that there is true single class to be allocated to each pixel, and the purpose of classification is to determine the class to which the pixel is allocated. Soft classifications seek to accommodate the mixed pixel, recognizing that the pixel areas may be composed of different land covers (*sensu* Fisher (1997)), or the ambiguity in class allocation and the degree to which the properties of the pixel match those of the defined classes. In both cases, the aim is to analyse the soft class probabilities or memberships rather than removing the uncertainty. This is increasingly of interest to the remote sensing community Silván-Cárdenas and Wang (2008); Khatami, Mountrakis, and Stehman (2017a); Zakeri and Mariethoz (2021).

This paper illustrated fuzzy approaches grounded in per-class membership differences and in the logic of fuzzy set intersection. The first approach computed measures of per class fuzzy certainty from the complement of the absolute difference between the observed and predicted fuzzy memberships. These were then interpolated using a GWR (Brunsdon, Stewart Fotheringham, and Charlton 1996) with the bandwidths determined optimally. Considerable variation in these was found across the classes, indicating different spatial variation in the relationship between predicted and observed fuzzy membership values. The fuzzy accuracy measures were generally high, indicating the similarity of the values collected in the field with those predicted by the remote sensing analysis. A second entirely novel GW fuzzy analysis transposed methods for quantifying fuzzy change from Fisher et al. (2006) into fuzzy certainty. It defined set operations for handling the intersection of different fuzzy membership values through the concept of bounded difference. These were used to construct a fuzzy correspondence matrix to capture the

magnitudes and directions of the uncertainty. The correspondence matrix diagonals were determined through the minimum operator, and the off-diagonals by transposing the concepts of bounded difference and fuzzy boundary (as used to for measures of fuzzy loss and fuzzy gain in Fisher et al. (2006)), to fuzzy commission and omission, and thus fuzzy user's and producers certainty. A GWR was used to interpolate fuzzy accuracy measures. The advantage of these approaches is that they provide more comprehensive measures of the certainty/uncertainty contained in the data. However, this requires a conceptualization of land cover objects captured through remote sensing as being inherently mixed and a rejection of the hard, crisp approach to classification.

A final analysis of the soft classification data was also undertaken by conceptualizing the probabilities and fuzzy memberships as compositional data (Tsutsumida et al. 2020; Zakeri and Mariethoz 2021). Methods for analysing differences in compositional data have been developed in domains that routinely work with count or measurement data, that are proportions of some whole. The key issue is the problem of using Euclidean distance in standard statistical measures of similarity, and a number of alternative metrics have been devised. Aitchison distance (Aitchison 1982) is one such measure and uses centred log-ratio transformation of the data. GW Aitchison distance was estimated in two ways, with different results: first, by calculating Aitchison distance from each of the 210 observations individually and then interpolating over the 1 km grid using a GWR; second by using a GW kernel to create a weighted data subset at each location from which a GW Aitchison distance was estimated, before interpolating over the grid. The differences reflect the different inputs used in each local model. The advantages of this approach are that it has a strong theoretical background in the biological and geological sciences, and there are a number of commonly used distance-based metrics for quantifying compositional data differences (e.g. chi-squared) as well those related to individual compositional elements (classes). These can easily be incorporated directly within either framework. The limitations of this approach are first that these are novel measures for the remote sensing community and may not be intuitively understood and they require a bandwidth to be defined, rather than determined optimally. Second, a more theoretically rigorous approach to handling zero values needs to be established. Future work will explore this approach and how to handle compositional data in more depth.

The key issues in all of the GW approaches relate to 1) whether the validation datasets contains sufficient observations, and 2) whether a GW kernel bandwidth can be optimized or has to be user specified. The Libyan validation data contains 210 observations. It is small, probably too small for robust inferences, but it is manageable and supports transparent GW model creation and evaluation. A related critical considerationfor example, when quantifying and interpolating class-specific measures, is whether the sample is sufficient across each class. This should probably be considered alongside the bandwidth – optimized or not (see below) – as this defines the number of observations used in each local metric. In practice, these issues are increasingly overcome with the use of high volume training and validation datasets. However, regardless, kernel size or bandwidth is a critical GW consideration as it defines the number of data points used in each local calculation and controls the degree of smoothing in the outputs, and thus the amount of spatial variation in them. Table 13 summarizes the ability to optimize bandwidths in the various GW approaches described in this paper. For those approaches where optimization is not

Table 13. A summary of the GW accuracy approaches, their underlying models and their ability to optimize GW kernel bandwidths.

GW accuracy model	Underlying GWR model	Bandwidth optimisation
GW-GLM	Generalized GWR	Yes
GWCM	None	No
GW Fuzzy certainty I	GWR	Yes
GW Fuzzy certainty II	GWR	Yes
GW Aitchison Distance	GWR	Yes

possible, bandwidth needs to be specified by the user, and justified in some way. In GWCM models, for example, the bandwidths were taken from the GW-GLM analysis, suggesting that a combination of GW accuracy evaluations may support these user choices.

6. Conclusions

There are opportunities to move beyond whole map (Openshaw 1996) accuracy measures in remote sensing, such as those reported in and derived from the standard correspondence matrix (e.g. overall, user's, producer's and kappa accuracies). The approaches described in this paper demonstrate how data collected as part of standard remote sensing training and validation exercise, can be used to generate spatially explicit measures of accuracy for hard classifications. It also extends these to the soft, probabilistic or fuzzy case. A critical observation is that the interpolated approaches to local accuracy reported here require validation data to be collected under an appropriate sampling framework following best practice recommendations (Olofsson et al. 2014; Stehman 2009) and may require considerable ground reference data for the local accuracy measures to be of value (Foody 2022). Whilst historically, the collection of soft validation data, usually through field survey in the manner reported in Comber et al. (2012), was complicated and required considerable effort and GPS precision, the situation is different now: with the increased availability of high resolution remote sensing data, it is quite common for training and validation data to be collected from remote sensing data of higher spatial resolution (and thus assumed quality) than the image to be classified. The result is that soft measures of probability or composition are now routinely available in training and validation data, opening up the opportunity to explore the classification uncertainties to a greater thematic and spatial depth. The methods described in this paper support these explicit spatial investigations of error, accuracy and uncertainty.

Acknowledgements

Disclosure statement

ORCID

Alexis Comber ⓘ http://orcid.org/0000-0002-3652-7846
Naru Tsutsumida ⓘ http://orcid.org/0000-0002-6333-0301

References

Aitchison, J. 1982. "The Statistical Analysis of Compositional Data." *Journal of the Royal Statistical Society: Series B (Methodological)* 44 (2): 139–160. https://doi.org/10.1111/j.2517-6161.1982.tb01195.x.

Aitchison, J. 2005. "A Concise Guide to Compositional Data Analysis." *In Compositional Data Analysis Workshop*.

Akaike, H. 1998. "Information Theory and an Extension of the Maximum Likelihood Principle." In *Selected Papers of Hirotugu Akaike*, 199–213. Springer New York. https://doi.org/10.1007/978-1-4612-1694-0_15.

Alshari, E. A., and B. W. Gawali. 2021. "Development of Classification System for LULC Using Remote Sensing and GIS." *Global Transitions Proceedings* 2 (1): 8–17. https://doi.org/10.1016/j.gltp.2021.01.002.

Atkinson, P. M. 2005. "Sub-Pixel Target Mapping from Soft-Classified, Remotely Sensed Imagery." *Photogrammetric Engineering & Remote Sensing* 71 (7): 839–846. https://doi.org/10.14358/PERS.71.7.839.

Bivand, R., D. Yu, T. Nakaya, M.-A. Garcia-Lopez, and M. R. Bivand. 2020. "Package 'Spgwr': spgwr: Geographically Weighted Regression." Accessed 30 September 2023. https://cran.r-project.org/web/packages/spgwr/spgwr.pdf

Brown, K. M., G. M. Foody, and P. M. Atkinson. 2009. "Estimating Per-Pixel Thematic Uncertainty in Remote Sensing Classifications." *International Journal of Remote Sensing* 30 (1): 209–229. https://doi.org/10.1080/01431160802290568.

Brunsdon, C., A. Stewart Fotheringham, and M. E. Charlton. 1996. "Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity." *Geographical Analysis* 28 (4): 281–298. https://doi.org/10.1111/j.1538-4632.1996.tb00936.x.

Comber, A. J. 2013. "Geographically Weighted Methods for Estimating Local Surfaces of Overall, User and Producer Accuracies." *Remote Sensing Letters* 4 (4): 373–380. https://doi.org/10.1080/2150704X.2012.736694.

Comber, A., C. Brunsdon, M. Charlton, and P. Harris. 2017. "Geographically Weighted Correspondence Matrices for Local Error Reporting and Change Analyses: Mapping the Spatial Distribution of Errors and Change." *Remote Sensing Letters* 8 (3): 234–243. https://doi.org/10.1080/2150704X.2016.1258126.

Comber, A., P. Fisher, C. Brunsdon, and A. Khmag. 2012. "Spatial Analysis of Remote Sensing Image Classification Accuracy." *Remote Sensing of Environment* 127:237–246. https://doi.org/10.1016/j.rse.2012.09.005.

Comber, A., and P. Harris. 2018. "Geographically Weighted Elastic Net Logistic Regression." *Journal of Geographical Systems* 20 (4): 317–341. https://doi.org/10.1007/s10109-018-0280-7.

Congalton, R. G. 1988. "Using Spatial Autocorrelation Analysis to Explore the Errors in Maps Generated from Remotely Sensed Data." *Photogrammetric Engineering and Remote Sensing* 54 (5): 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. https://doi.org/10.1016/0034-4257(91)90048-B.

Ebrahimy, H., B. Mirbagheri, A. Akbar Matkan, and M. Azadbakht. 2021. "Per-Pixel Land Cover Accuracy Prediction: A Random Forest-Based Method with Limited Reference Sample Data." *ISPRS Journal of Photogrammetry and Remote Sensing* 172:17–27. https://doi.org/10.1016/j.isprsjprs.2020.11.024.

Fisher, P. 1997. "The Pixel: A Snare and a Delusion." *International Journal of Remote Sensing* 18 (3): 679–685. https://doi.org/10.1080/014311697219015.

Fisher, P., C. Arnot, R. Wadsworth, and J. Wellens. 2006. "Detecting Change in Vague Interpretations of Landscapes." *Ecological Informatics* 1 (2): 163–178. https://doi.org/10.1016/j.ecoinf.2006.02.002.

Fisher, P. F., and S. Pathirana. 1990. "The Evaluation of Fuzzy Membership of Land Cover Classes in the Suburban Zone." *Remote Sensing of Environment* 34 (2): 121–132. https://doi.org/10.1016/0034-4257(90)90103-S.

Foody, G. M. 2020. "Explaining the Unsuitability of the Kappa Coefficient in the Assessment and Comparison of the Accuracy of Thematic Maps Obtained by Image Classification." *Remote Sensing of Environment* 239:111630. https://doi.org/10.1016/j.rse.2019.111630.

Foody, G. M. 2022. "Global and Local Assessment of Image Classification Quality on an Overall and Per-Class Basis without Ground Reference Data." *Remote Sensing* 14 (21): 5380. https://doi.org/10.3390/rs14215380.

Fotheringham, A. S., C. Brunsdon, and M. Charlton. 2002. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. London: John Wiley & Sons.

Gollini, I., B. Lu, M. Charlton, C. Brunsdon, and P. Harris. 2015. "GWmodel: An R Package for Exploring Spatial Heterogeneity Using Geographically Weighted Models." *Journal of Statistical Software* 63 (i17). https://doi.org/10.18637/jss.v063.i17

Khatami, R., G. Mountrakis, and S. V. Stehman. 2017a. "Mapping Per-Pixel Predicted Accuracy of Classified Remote Sensing Images." *Remote Sensing of Environment* 191:156–167. https://doi.org/10.1016/j.rse.2017.01.025.

Khatami, R., G. Mountrakis, and S. V. Stehman. 2017b. "Predicting Individual Pixel Error in Remote Sensing Soft Classification." *Remote Sensing of Environment* 199:401–414. https://doi.org/10.1016/j.rse.2017.07.028.

Nakaya, T. 2015. "Geographically Weighted Generalised Linear Modelling." *Geocomputation: A Practical Primer* 217 (20): 201–220.

Nakaya, T., A. S. Fotheringham, C. Brunsdon, and M. Charlton. 2005. "Geographically Weighted Poisson Regression for Disease Association Mapping." *Statistics in Medicine* 24 (17): 2695–2717. https://doi.org/10.1002/sim.2129 .

Olofsson, P., G. M. Foody, M. Herold, S. V. Stehman, C. E. Woodcock, and M. A. Wulder. 2014. "Good Practices for Estimating Area and Assessing Accuracy of Land Change." *Remote Sensing of Environment* 148:42–57. https://doi.org/10.1016/j.rse.2014.02.015 .

Openshaw, S. 1996. "Developing GIS-Relevant Zone-Based Spatial Analysis Methods." In Longley, P, and Batty, M. (Eds.), *Spatial Analysis: Modelling in a GIS Environment* (pp. 55–73). Cambridge: GeoInformation International.

Pontius, R. G., Jr, and M. Millones. 2011. "Death to Kappa: Birth of Quantity Disagreement and Allocation Disagreement for Accuracy Assessment." *International Journal of Remote Sensing* 32 (15): 4407–4429. https://doi.org/10.1080/01431161.2011.552923.

Pontius, R. G., Jr, and A. Santacruz. 2014. "Quantity, Exchange, and Shift Components of Difference in a Square Contingency Table." *International Journal of Remote Sensing* 35 (21): 7543–7554. https://doi.org/10.1080/2150704X.2014.969814.

Pontius, R. G., Jr, and A. Santacruz. 2015. "Package'diffeR': Metrics of Difference for Comparing Pairs of Maps." Accessed 30 September 2023. https://cran.r-project.org/web/packages/diffeR/diffeR.pdf

Silván-Cárdenas, J. L., and L. Wang. 2008. "Sub-Pixel Confusion–Uncertainty Matrix for Assessing Soft Classifications." *Remote Sensing of Environment* 112 (3): 1081–1095. https://doi.org/10.1016/j.rse.2007.07.017.

Stehman, S. V. 2009. "Sampling Designs for Accuracy Assessment of Land Cover." *International Journal of Remote Sensing* 30 (20): 5243–5272. https://doi.org/10.1080/01431160903131000.

Templ, M., K. Hron, and P. Filzmoser. 2011. "Package 'robCompositions': Compositional Data Analysis." Accessed 30 September 2023. https://cran.r-project.org/web/packages/robCompositions/robCompositions.pdf

Tsutsumida, N., P. Rodrguez-Veiga, P. Harris, H. Balzter, and A. Comber. 2019. "Investigating Spatial Error Structures in Continuous Raster Data." *International Journal of Applied Earth Observation and Geoinformation* 74:259–268. https://doi.org/10.1016/j.jag.2018.09.020.

Tsutsumida, N., T. Yoshida, D. Murakami, and T. Nakaya. 2020. "A Geographically Weighted Total Composite Error Analysis for Soft Classification." In *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*, 874–876. Waikoloa, HI, USA: IEEE. https://doi.org/10.1109/IGARSS39084.2020.9323939

Wang, F. 1990. "Fuzzy Supervised Classification of Remote Sensing Images." *IEEE Transactions on Geoscience and Remote Sensing* 28 (2): 194–201. https://doi.org/10.1109/36.46698.

Wang, Q., W. Shi, and P. M. Atkinson. 2014. "Sub-Pixel Mapping of Remote Sensing Images Based on Radial Basis Function Interpolation." *ISPRS Journal of Photogrammetry and Remote Sensing* 92:1–15. https://doi.org/10.1016/j.isprsjprs.2014.02.012.

Zadeh, L. A. 1965. "Fuzzy Sets." *Information & Control* 8 (3): 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X.

Zakeri, F., and G. Mariethoz. 2021. "A Review of Geostatistical Simulation Models Applied to Satellite Remote Sensing: Methods and Applications." *Remote Sensing of Environment* 259:112381. https://doi.org/10.1016/j.rse.2021.112381.