CHAPTER 5

GEOGRAPHICALLY WEIGHTED METHOD FOR ANALYSIS OF SPATIAL VARIATION IN CLASSIFICATION ACCURACY

5.1 INTRODUCTION

In recent years, remote sensing images have been employed broadly to extract thematic information through digital image classification techniques. Assessing the accuracy of regional to global scale thematic maps derived from remote sensing is recognized as an essential requirement to support most of the mapping projects, scientific applications and policy decisions (Foody, 2002; Strahler et al., 2006). In remote sensing, the confusion matrix and its associated measures, such as OA, UA and PA have become the conventional paradigm for reporting the mapping accuracy (Congalton and Green, 1999). Descriptions of accuracy can help to appraise the uncertainties coupled with thematic data or to choose between thematic datasets when there is the immense accessibility of data with different thematic or spatial characteristics (See and Fritz, 2006). Therefore, accuracy is one of the leading features of any remote sensing data product. However, there are some restrictions associated with the paradigm of the confusion matrix. The global estimates of accuracy metrics are inadequate to express the overall quality of thematic maps, as these do not reflect the spatial distribution of errors over the image. It may not be appropriate for local sub-regions, where error rates may be much larger or smaller than the global estimates (McGwire and Fisher, 2001; Foody, 2005). Several studies have reported different types of spatially distributed errors and methods to conquer this problem (Foody 2005; Steele et al., 1998; Riemann et al., 2010). The existing validation and accuracy assessment methods have been largely disregarded the advances supported by such methods by remote sensing community.

Therefore, the quality of thematic maps derived from remotely sensed data needs to be budgeted and requires improved methods or tools for estimating and describing the spatial distribution of errors in landscape mapping. A phenomenon differs across a landscape because of spatial non-stationarity or heterogeneity. This spatial variability restricts to employ any conventional global regression technique which assumes that the observations are independent of the spatial location. The use of conventional regression techniques e.g. Ordinary Least Square Regression (OLSR) lead to erroneous conclusions in spatial analysis and produces spatially autocorrelated residuals (LeSage and Pace, 2001). Alternatively, several local regression techniques have been suggested to address the challenges caused by spatial non-stationarity. One of the best recognized approaches for spatial regression is the GWR), a statistical technique that explicitly deals with spatial non-stationarity (Brunsdon et al., 1996; Fotheringham et al., 2002). GWR is a local regression technique allows for the computation of relationships among variables varying over geographical space (Fotheringham et al., 2002). It computes the local estimates of the regression coefficients for a moving geographic window or kernel at every location. Unlike conventional regression techniques which encapsulate the global relationship among the variables in a single regression equation, GWR creates spatially varying data for the relationships among variables. In several studies the better performance of GWR has been reported for various applications (Brunsdon et al., 1996; Leung et al., 2000; Zhang and Gove, 2005).

The focus of our attention on GWR is motivated by numerous studies which have demonstrated its potential in the investigation of spatially varying relationships, including climatology (Brunsdon et al., 2001), health (Lin and Wen, 2011; Ehlkes et al., 2014), real estate management (Lu et al., 2014) and urban studies (Faisal and Shaker, 2017). GWR is

explored to investigate the spatial variation of the relationship between land cover and population change (Maimaitijiang et al., 2015). GWR can also be applied in combination with linear, logistic and poisson regression techniques for various applications. Geographically weighted poisson regression model is used for disease mapping (Ehlkes et al., 2014; Nakaya et al., 2005). Lesive et al., (2016) explored geographically weighted logistic regression technique for the comparison of data fusion of different land cover products derived from remote sensing image.

However, the application of GWR integrated with logistic regression for generating spatially varying accuracies of a heterogeneous landscape mapping using high resolution remote sensing image is still limited. This research aims to explore geographically weighted logistic regression method for estimating spatial variation in accuracy measures and compare it with a conventional global OLSR technique. It may be used in different disciplines for instance economics, environment, social and earth sciences, where improved understanding about the local behavior of parameter relations is needed. The consequences of present work may also be helpful for addressing long-standing gaps in the analysis and description of spatially explicit accuracy of thematic information of a heterogeneous landscape.

5.2 STUDY SITE AND MATERIALS

The study site for this work, extending from 82° 54′ 30″ to 83° 02′ 30″ E, and from 25° 13′ 08″ to 25° 20′ 43″ N, covering a total area of 25327 ha. A remote sensing image acquired on 6 April 2013 from LISS-IV sensor satellite with high spatial resolution of 5.8 m was used in this study. It was classified into six major LULC classes such as agricultural land, vegetation, fallow land, built up, water bodies and sand according to the landscape of study site using random forest (RF) classifier (Breiman, 2001). The mapping and regression

analyses were implemented in an open source software R version 3.4.0 (http://cran.rproject.org). The precise geolocation of validation data set was collected at 551 locations by random sampling method with the help of a handheld GPS receiver (Trimble Juno 3B). The study area with the sample point locations are shown in Figure 5.1 as viewed on FCC of LISS-IV image. The OA was calculated from the diagonal and off-diagonal components while, UA and PA were calculated using the row and column totals in the confusion matrix.

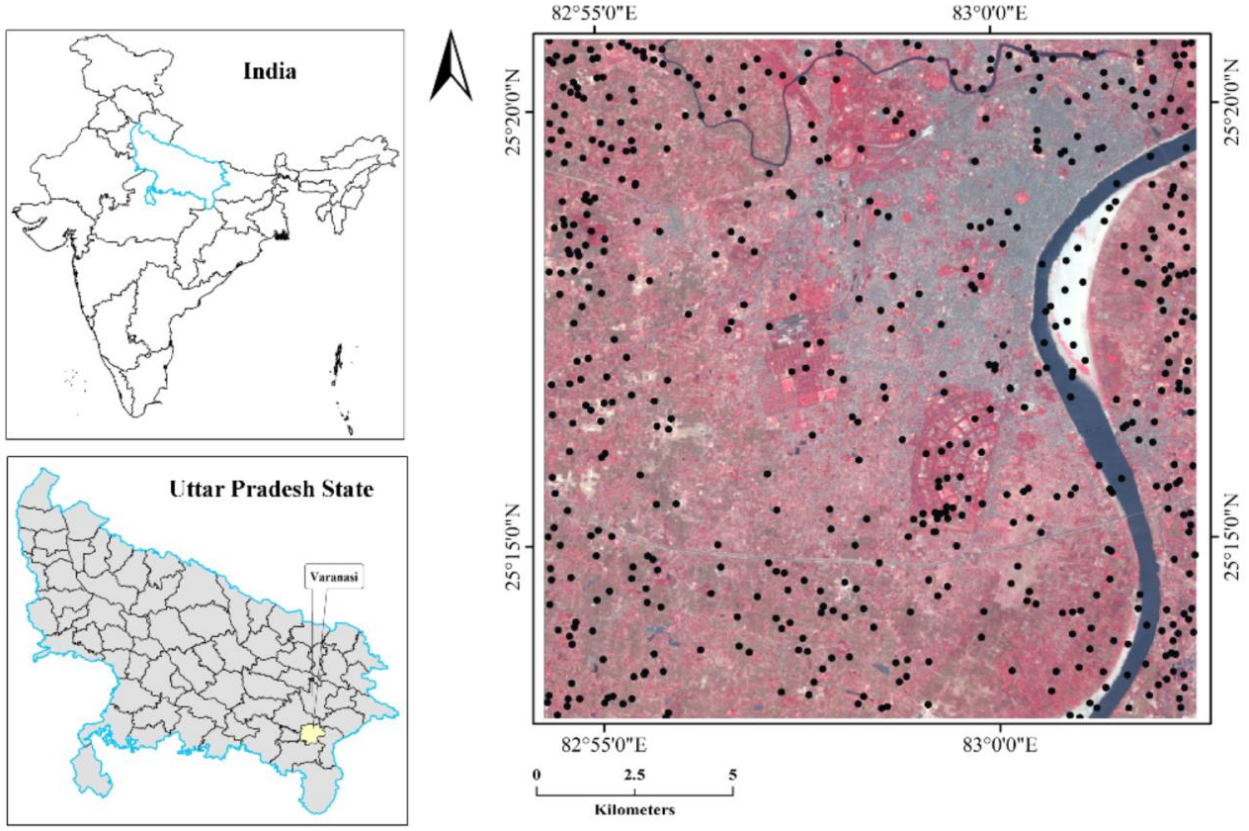

Figure 5.1 Location of the study site with the validation points (●) as viewed on FCC of LISS-IV sensor image

5.3 METHODOLOGY

Regression analysis is the most commonly used statistics to examine and explore the spatial relationships among the variables. In spatial data analysis, several regression

techniques have been described and formulated over the years. The brief description of regression techniques used in this study is given as:

5.3.1 OLSR

OLSR is a generalized linear regression technique. It may be applied to single or multiple explanatory variables. This regression technique estimates the coefficients by using ordinary least squares.

Given a set of $n\ (k = 1,2, \ldots, n)$ observations on $p\ (g = 1,2, \ldots, p)$ independent variables X, and a dependent variable Y, the relationship between Y and X can be regressed using OLS as

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (5.1)$$

where $i = 1 \ldots . n$

The error term or residual ε_i and parameters β_0, β_1 are to be estimated. The OLS estimate of β is obtained by Equation given as:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (5.2)$$

where $\hat{\beta}$ is the vector of estimated parameters, X is the matrix of independent variables preceded by a vector of 1s, Y is the vector of n observed values of dependent variables, $(X^T X)^{-1}$ is the inverse of variance-covariance matrix, T is transpose of a matrix.

Since the weights can also be included in OLS estimator and placed in the diagonal of a square matrix. The weight of i^{th} observation is w_i and W is the matrix with w_i weights on its diagonal. The estimator with the weights are shown in Equation (5.3)

$$\hat{\beta} = (X^T W X)^{-1} X^T Y \quad (5.3)$$

5.3.2 GWR

GWR is a spatial statistical technique which extends the framework of conventional regression statistics by revising a globally defined model as a locally estimated model. It enables meaningful analysis in modeling spatially heterogeneous processes (Fotheringham et al., 2002). In contrast to global regression, the coefficients in GWR are functions of varying spatial location. The general form of a basic GWR model is given by (Fotheringham et al., 2002) and can be written as:

$$y_i = \beta_{i0} + \sum_{k=1}^{m} \beta_{ik} x_{ik} + \varepsilon_i \qquad (5.4)$$

where, y_i is the dependent variable at location i; x_{ik} is the independent variable k at location i; m is the number of independent variables; β_{i0} is the intercept parameter at location i; β_{ik} is the is the local regression coefficient for the kth independent variable at location i; and ε_i is the random error at location i. The coefficients in GWR differ continuously across the study landscape. A set of coefficients can be estimated at any location by the given dependent variable and one or more independent variables which have been measured at the spatial location with known coordinates. For a given coordinates (u_i, v_i) at the location i, GWR Equation can also be expressed as:

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{m} \beta_k(u_i v_i) x_{ik} + \varepsilon_i \qquad (5.5)$$

GWR determines the implicit relationships around each regression point i, where all the regression coefficients is estimated by weighted least squares approach, for which the matrix expression is given as:

$$\hat{\beta}_i = (X^T W(u_i v_i) X)^{-1} X^T W(u_i v_i) Y \quad (5.6)$$

where X is the matrix of independent variables; $\hat{\beta}_i = (\beta_{i0}, \ldots . \beta_{im})^T$ is the vector of m+1 local regression coefficients; and $W(u_i, v_i)$ is the n by n weighting matrix whose diagonal elements are indicating geographical weights of each observation for regression point and off-diagonal elements are zero. It can be expressed as:

$$\begin{bmatrix} w_{i1} & 0 & \ldots & 0 \\ 0 & w_{i2} & \ldots & 0 \\ \vdots & \vdots & \vdots & \\ 0 & 0 & \ldots & w_{i3} \end{bmatrix} \quad (5.7)$$

The weighting system W_i is calculated via a kernel function from proximities between regression point i and the N data points around it. In this study, a Gaussian kernel is specified and can be given as:

$$\text{Gaussian} : w_{ij} = \exp\left[-\frac{1}{2}\left(\frac{d_{ij}}{b}\right)^2\right] \quad (5.8)$$

where d_{ij} is the distance between regression point i and observation point j and b is the kernel bandwidth.

An optimal bandwidth can be selected normally using goodness-of-fit measure. In this work, the cross-validation (CV) approach (Bowman, 1984) was used in which bandwidth is selected by minimizing the CV score. The CV score is calculated by

$$CV = \sum_{i=1}^{n} (y_i - \hat{y}_{i \neq i})^2 \quad (5.9)$$

where n is the number of observations, and observation i is omitted from the calculation so that in areas of sparse observations the model is not calibrated solely on i.

5.3.3 GWLR

Since, the presence or absence of a specific land cover category is a binary outcome (1/0, Yes/No, True/False). Therefore, a powerful analytical technique namely logistic regression or logit model is used (Peng et al., 2002). The logit function is defined by

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) \tag{5.10}$$

Logistic models provide a probability ranging from 0 to 1, representing the correct prediction of a land cover category. When it is used in combination with GWR, it allows for the spatially varying local estimation of correctly classified and incorrectly classified land cover categories. When the response variable is binary, GWR should be applied via geographically weighted logistic regression (GWLR). It is applied to examine how the classification accuracy varied across geographical space. GWLR is the geographically weighted extension to the logistic regression model. It is similar in form to ordinary regression, but geographically weighted techniques use a moving window or kernel under which local regression models are computed at locations all over the study region. The GWR Equation (5.4) can be extended to GLWR with the help of a logit function in the following way:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0(u_i, v_i) + \sum_{k=1}^{m} \beta_k(u_i v_i) X_{ik} + \varepsilon_i \tag{5.11}$$

or it may be written in the following way:

$$p_i = \frac{e^{\beta_0(u_i,v_i)+\sum_{k=1}^{m}\beta_k(u_i v_i)x_{ik}}}{1+e^{\beta_0(u_i,v_i)+\sum_{k=1}^{m}\beta_k(u_i v_i)x_{ik}}} \tag{5.12}$$

where p_i is probability of prediction at location i, and other terms are same as in Equation (5.4).

Spatially varying accuracy measures were estimated by applying GWLR which compares the validation and classified data. GWLR was applied to model the relationship between the random forest classifier based class and the class from the validation data by logistic transformation given as:

$$\ln\left(\frac{P(O_i = 1)}{1 - P(O_i = 1)}\right) = \beta_{0(u_i,v_i)} \quad (5.13)$$

$$\ln\left(\frac{P(y_i = 1)}{1 - P(y_i = 1)}\right) = \beta_{0(u_i,v_i)} + \beta_1 x_{1(u_i,v_i)} \quad (5.14)$$

$$\ln\left(\frac{P(x_i = 1)}{1 - P(x_i = 1)}\right) = \beta_{0(u_i,v_i)} + \beta_1 y_{1(u_i,v_i)} \quad (5.15)$$

where $P(O_i = 1)$, $P(y_i = 1)$ and $P(x_i = 1)$ are the probabilities of OA, UA and PA respectively at location i. β_0 is the intercept, β_1 is the slope and (u_i, v_i) is the two-dimensional co-ordinates representing the location of i. The coefficients in GWLR are permitted to vary across the two-dimensional geographical space characterized by the coordinates (u, v) within the study region.

5.3.4 Comparison of OLSR and GWLR

To compare the model performance between GWLR and OLSR the statistical parameters namely adjusted squared correlation coefficient (R^2), Residual sum of squares (RSS) and Akaike Information Criterion (AIC) were employed. The R^2 measures the goodness of fit and varies from 0 to 1. It is likely to be higher when more variance is explained by the dependent variable. The lower AIC value describes the stronger ability of the regression technique to reflect reality. The RSS is used to measure the quantity of variance in the data set that is not explained by the regression technique. It computes the amount of error remaining between the regression function and data set. A small RSS

signifies a robust fit of the model and explains a larger amount of the data. Better performance of regression technique is indicated by getting higher R^2 value with lower RSS and AIC values.

5.4 RESULTS

A standard confusion matrix was constructed to calculate UA, PA and OA using validation dataset (Table 5.1). At the same time, it is evident that some classes are classified more presumably than the others. Also, the table does not provide information about the spatial distribution of errors associated with different landscape classes. The GWLR was applied to examine the spatial variation in the relationship between data classified from the remote sensing image and the reference data collected through field survey. Table 5.2 illustrates the spatial distribution of accuracy measures (UA, PA and OA) for each landscape category in terms of minimum, median, maximum, 1st and 3rd quartiles along with the inter-quartile range (IQR). The IQR is described as a representative metric of the overall spatial variation in accuracy measures.

The larger IQR values signify the greater spatial variation (Comber et al., 2012). It also exhibits the extent to which the observed reference data (ground truth) are inferred by the predicted data (from remote sensing image), and the variation in GWLR method. The IQR is described as a representative metric of the overall spatial variation in accuracy measures. The IQR values of UA, PA and OA for various landscape categories are listed in Table 5.2.

Table 5.1 Confusion matrix comparing reference and classified landscape categories, showing 87.48% overall agreement

	Reference							
Classified	Sand	Vegetation	Water bodies	Agricultural land	Fallow land	Built up	Row total	UA
Sand	19	0	0	0	5	10	34	0.559
Vegetation	0	100	1	9	0	0	110	0.909
Water bodies	0	0	91	0	0	0	91	1.000
Agricultural land	0	8	0	101	0	0	109	0.927
Fallow land	0	0	0	0	78	14	92	0.848
Built up	1	1	8	5	7	93	115	0.809
Column total	20	109	100	115	90	117	551	
PA	0.950	0.917	0.910	0.878	0.867	0.795		
OA				0.875				
Kc				0.846				

Table 5.2 Summary of the variation in GWLR representing UA, PA and OA for various landscape categories. IQR indicates the variability in probability among pixels. 1st and 3rd quartiles represent the 25th and 75th percentile probabilities

Categories		Min.	1St Quartile	Median	Mean	3rd Quartile	Max.	IQR
Agricultural land	UA	0.015	0.800	0.999	0.851	1.000	1.000	0.200
	PA	0.533	0.793	0.896	0.875	0.972	0.998	0.179
	OA	0.004	0.644	0.807	0.751	0.976	1.000	0.333
Vegetation	UA	0.627	0.872	0.955	0.921	0.995	1.000	0.124
	PA	0.377	0.882	0.995	0.922	1.000	1.000	0.119
	OA	0.343	0.780	0.911	0.859	0.979	1.000	0.199
Fallow land	UA	0.041	0.802	0.997	0.874	1.000	1.000	0.198
	PA	0.381	0.779	0.879	0.857	0.967	1.000	0.188
	OA	0.017	0.636	0.874	0.782	0.995	1.000	0.359

	UA	0.271	0.741	0.880	0.845	0.975	1.000	0.234
Built up	PA	0.095	0.724	0.861	0.823	0.980	1.000	0.256
	OA	0.069	0.580	0.717	0.724	0.918	0.998	0.338
	UA	1.000	1.000	1.000	1.000	1.000	1.000	0.000
Water bodies	PA	0.530	0.866	0.973	0.923	0.999	1.000	0.133
	OA	0.530	0.866	0.977	0.923	0.999	1.000	0.133
	UA	0.000	0.002	0.374	0.381	0.737	0.932	0.735
Sand	PA	0.494	0.990	1.000	0.944	1.000	1.000	0.010
	OA	0.000	0.000	0.008	0.308	0.721	1.000	0.721

For agricultural land, OA was found to vary from 0.004 to 1.000 in different parts of the study area in comparison to a global figure of 0.875. While, UA was found to vary from 0.015 to 1.000 as compared to a global figure of 0.927 and PA was found to vary from 0.533 to 0.998 as compared to a global figure of 0.878 across the study area. For vegetation, OA was found to vary from 0.343 to 1.000 in different parts of the study area in comparison to a global figure of 0.875. While, UA was found to vary from 0.627 to 1.000 as compared to a global figure of 0.909 and PA was found to vary from 0.377 to 1.000 compared to a global figure of 0.917 across the study area. For fallow land, OA was found to vary from 0.017 to 1.000 in different parts of the study area in comparison to a global figure of 0.875. While, UA was found to vary from 0.041 to 1.000 compared to a global figure of 0.848 and PA was found to vary from 0.381 to 1.000 compared to a global figure of 0.867 across the study area. For built up, OA was found to vary from 0.069 to 0.998 in different parts of the study area in comparison to a global figure of 0.875. While, UA was found to vary from 0.271 to 1.000 compared to a global figure of 0.809 and PA was found to vary from 0.095 to 1.000 compared to a global figure of 0.795 across the study area. For water bodies, OA was found to vary from 0.530 to 1.000 in different parts of the study area in comparison to a global

figure of 0.875. While, UA was found to vary from 1.000 to 1.000 compared to a global figure of 1.000 and PA was found to vary from 0.530 to 1.000 compared to a global figure of 0.910 across the study area. For sand, OA was found to vary from 0.000 to 1.000 in different parts of the study area in comparison to a global figure of 0.875. While, UA was found to vary from 0.000 to 0.932 compared to a global figure of 0.559 and PA was found to vary from 0.494 to 0.010 compared to a global figure of 0.950 across the study area. In the context of UA, the sand class exhibited larger while, the water bodies exhibited smaller spatial variation. In the context of PA, the built up class exhibited larger while, the sand class exhibited smaller spatial variation. It is remarkable to note that the distribution of UA estimates reveals much larger variation followed by the moderate variation in the distribution of OA estimates, whereas smaller variation was observed in the distribution of PA estimates. The visual representation of spatial variations in UA, PA and OA are shown in Figures 5.2, 5.3 and 5.4 respectively.

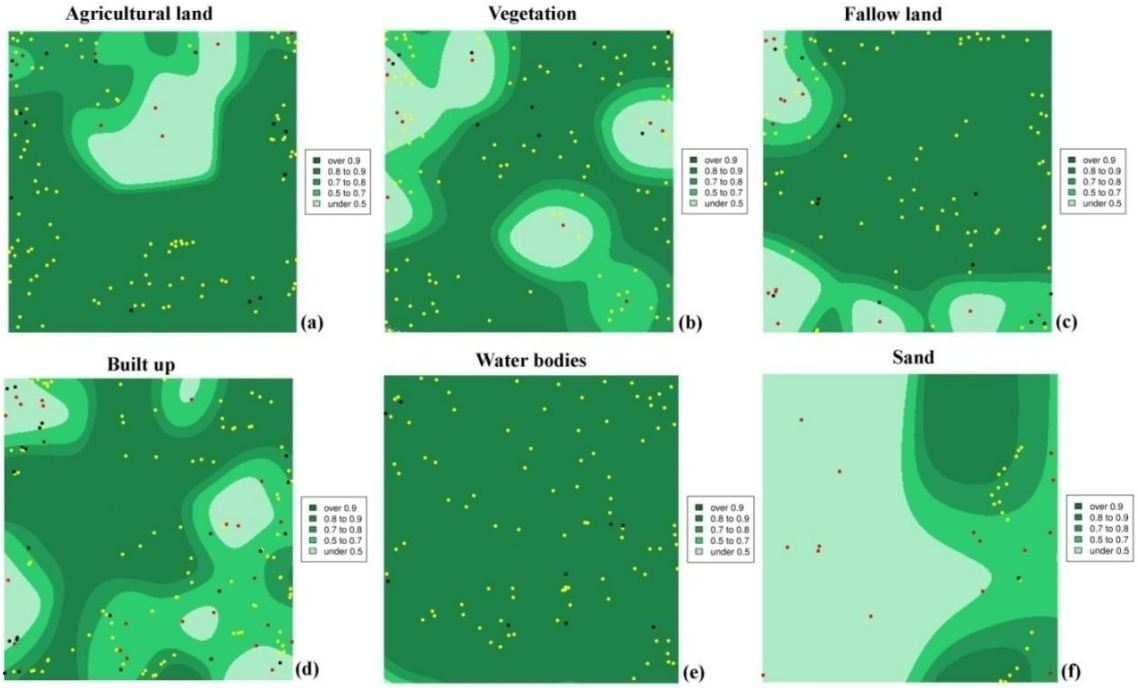

Figure 5.2 GWLR based maps of spatial variation in UA of landscape categories

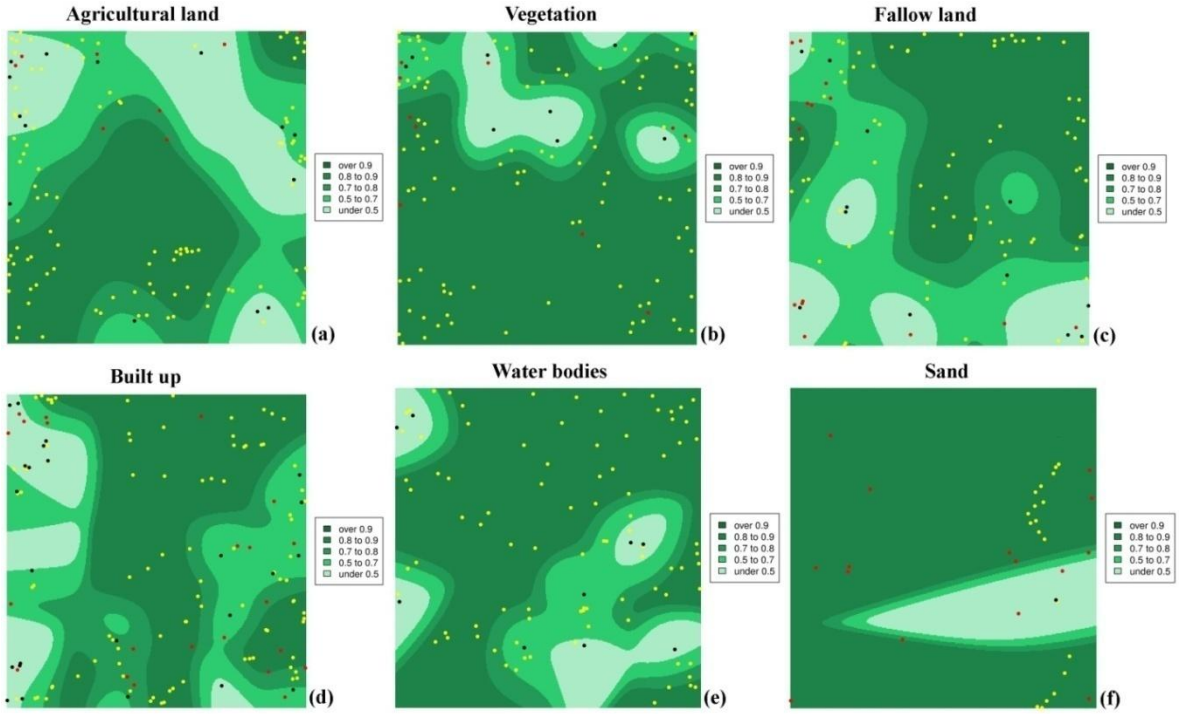

Figure 5.3 GWLR based maps of spatial variation in PA of landscape categories

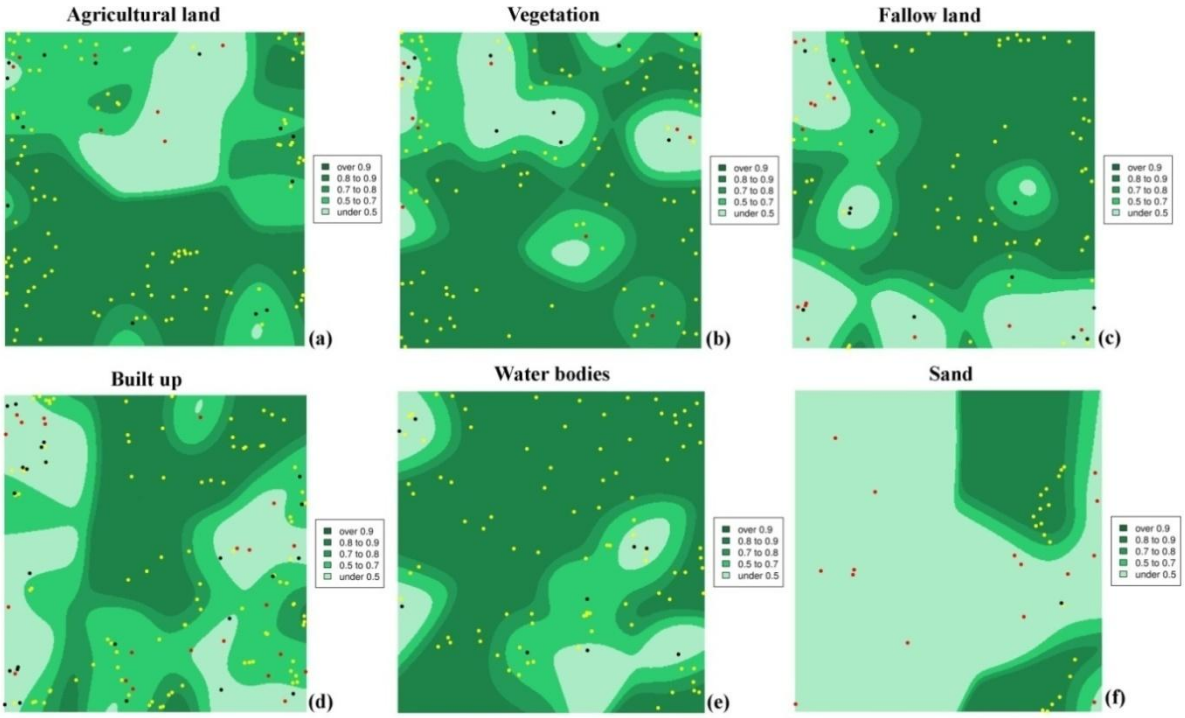

Figure 5.4 GWLR based maps of spatial variation in OA of landscape categories

Note: The legend values in figures represent the probability of the presence of the correct category. Yellow circles represent locations correctly classified. Red circles represent locations incorrectly classified as given category (commission error). Black circles represent locations incorrectly classified to other categories (omission error).

For major landscape category i.e. agricultural land, the UA and PA are higher in the southern and lower towards the northern parts of the study area (Figures 5.2 (a) and 5.3 (a)). Figure 5.3 (a) shows that there is higher variation in the estimates of PA of agricultural land, but there is a trend of marginally higher variation towards the northern part of the study area. Figure 5.4 (a) shows that the OA for agricultural land is found to be higher in the southern and lower towards the northern parts of the study area. Figure 5.2 (d) shows that the UA for other major landscape category built up is higher in the middle and lower in the south-east parts of the study area. Figure 5.3 (d) shows that the PA for built up is higher in the middle and lower in the western parts of the study area. Figure 5.4 (d) shows that the OA for built up category is found to be higher in the middle and lower in the south-east and western parts of the study area. Other landscape categories are also showing spatial variation in the accuracy measures (Figure 5.2-5.4). The UA is the estimation of the probability that a classified pixel correctly represents the categories on the ground. It shows the commission error (inclusion) and for the user of the map it signifies the probability of locating that class on the ground correctly. Here, Figure 5.2 (a) would suggest less confidence in areas mapped as agricultural land actually being that class on the ground. Also, Figure 5.2 (d) would suggest less confidence in areas mapped as built up actually being that class on the ground. The PA is the estimation of the probability that a reference pixel is correctly identified in the classified data. It shows the omission error (exclusion) and for the producer of the map, it signifies the

probability that the classes of interest are omitted from the classified data. Therefore, it is convinced that most of the agricultural land that exists has been mapped with low omission error. However, there are high levels of commission errors in the northern part of the study area as well. In the case of built up category, it is convinced that most of the class of interest exists has been mapped with low omission error. However, there are high levels of commission errors in the south-east part of the study area as well.

The performance of estimating spatial variation in overall accuracies was compared using OLSR and GWLR for different landscape categories. As can be seen from Table 5.3, the GWLR produced relatively lower AIC and RSS values as compared to OLSR, which indicated a better fit to the observed data. While higher R^2 values for GWLR in comparison to OLSR showed that more variance is explained for the dependent variable. The detailed information is shown in Table 5.3. Overall, the performance of GWLR technique was found to be better in comparison to OLSR for estimation of spatially varying accuracy of landscape mapping.

Table 5.3 Comparison of GWLR and OLSR techniques for estimating spatial variation in overall accuracies of various landscape categories

Landscape categories	Model	GWLR			OLSR		
	Parameters	Adjusted R^2	RSS	AIC	Adjusted R^2	RSS	AIC
	Agricultural land	0.241	11.81	64.81	0.077	18.07	117.12
	Vegetation	0.179	10.48	52.55	0.097	15.97	102.68
	Fallow land	0.450	7.64	27.62	0.011	19.50	125.05
	Built up	0.197	19.52	125.57	0.032	30.78	188.89
	Water bodies	0.155	6.08	7.71	0.049	8.19	37.56
	Sand	0.647	1.65	-3.68	0.075	8.69	54.55

5.5 DISCUSSION

The confusion matrix is a convenient way to summarize errors of landscape mapping, but it is aspatial in nature. In landscape mapping with comparatively high accuracy, it may not be necessarily applicable to map spatially varying local accuracy. However, a spatial representation could be beneficial for landscape mapping in low accuracy or problematic categories having poor accuracy. Geographical analyses can be used to understand how and why processes vary spatially. The locational attributes of data are used explicitly to recognize local variations in relationships. It is not similar as analyzing spatial data in itself under the hypothesis that it represents a spatial analysis as the data are spatial in nature. In remote sensing community, it is confusing to implement spatially explicit approach because error cluster of land cover features are well identified for a variety of well-reported causes. Therefore, it is needed to evaluate the local error frequently by familiar visual and qualitative techniques (Friedl et al., 2002). This work demonstrates how the logistic regression can be used to produce probabilities of accuracy measures, and the ability of their geographically weighted extension to generate spatial distributions that illustrate the variation of these probabilities. The remote sensing society is perhaps well known with the concepts of OA, PA and UA and the method proposed here may better reflect their needs.

This study does not attempt to overcome all of the limitations associated with the confusion matrix. Rather, this study investigated spatially explicit methods for describing accuracy using geographically weighted methods to identify spatially varying relationship between classified and reference data. The capability to estimate spatially explicit accuracy measures and errors from data collected as part of standard validation work out, recommends that the maps of the distribution of accuracy could assist confusion matrices. Geographically

weighted methods can be used to produce spatially explicit outputs which point out the potential advantages of incorporating the results of any validation exercise along with the remote sensing data based land cover product. Finally, the method proposed in this study addresses one of the fundamental beliefs of geographical analyses that the process under examination varies over space continuously.

5.6 CONCLUSION

The geographically weighted technique was used to describe the spatial variation of accuracy by random forest classification using high resolution remote sensing data. It addresses major concerns in the analysis and description of accuracy and errors associated with heterogeneous landscape mapping. It also provides a better understanding of non-stationarity in landscape errors which frequently vary by analyzing its spatial distribution. This work shows how logistic regression can be applied to produce probabilities such as UA, PA and OA and the ability of its geographically weighted extensions to generate spatial distributions describing the variation of these probabilities. This study also compared the performance of GWLR with conventional OLSR technique. The investigation showed better performance of GWLR in estimating spatially varying accuracy measures compared to OLSR. Finally, spatially explicit accuracy measures are more informative and precise because they are spatial and offer better support for assessments of data accuracy than the confusion matrix based global measures. The results of this work suggest that there is a need to reconsider the tenets of accuracy and errors associated with remote sensing.