Centre for Research on Settlements and Urbanism

Journal of Settlements and Spatial Planning

Journal homepage: http://jssp.reviste.ubbcluj.ro

Accounting for Spatial Non-Stationarity to Estimate Population Distribution Using Land Use/Cover. Case Study: the Lake Naivasha Basin, Kenya

Dawit W. MULATU[1], Anne VAN DER VEEN[1], Robert BECHT[1], Pieter R. VAN OEL[1], Desta J. BEKALO[1]

[1] *University of Twente, ITC-Faculty of Geo-Information Science and Earth Observation, Enschede, THE NETHERLANDS*

E-mail: mulatu25556@itc.nl, veen@itc.nl, becht@itc.nl, oel@itc.nl, bekalo05012@itc.nl

K e y w o r d s: *land use/cover, remotely-sensed data, population density, non-stationarity, Lake Naivasha basin, geographically weighted regression (GWR), ordinary least squares (OLS)*

A B S T R A C T

Remotely-sensed data can be used to overcome deficiencies in data availability in poorly monitored regions. Reliable estimates of human population densities at different spatial levels are often lacking in developing countries. This study explores the applicability of a geographically-weighted regression (GWR) model for estimating population densities in rural Africa using land use/cover data that have been derived from remote-sensing while accounting for spatial non-stationarity. This study was conducted for the Lake Naivasha basin in Kenya where population pressure, intense land utilization in the catchment and informal settlements in Naivasha town due to lucrative economic activities are the major challenges of the basin socio-ecological system. The results of this study show that using a GWR model for taking into account the spatially-varying relationship between specific land use/cover classes and population significantly improves population estimates and handles the spatial non-stationarity that could not be addressed by global ordinary least squares (OLS) model. The result revealed that the parameter estimates (coefficients) for grassland and cropland use/cover have a significant spatially varying relationship with population and exhibit locally different signs, which would have gone undetected by a global model. Consequently, this study indicates that incorporating spatial non-stationarity can significantly improve population density estimates for rural Africa based on remotely-sensed data.

1. INTRODUCTION

The relationship between human population distribution and land use/cover change is much debated [1]. The spatial distribution of human population on the land surface is considered as a fundamental determinant of land use impacts on natural ecosystems [2]. In developing countries, excessive high rural human population density is a frequent concern in terms of overpopulation and pressure on environmental carrying capacity. But when population density gets too low, it also has adverse impacts on rural areas [3]. On a large spatial scale, human presence is positively related to biodiversity suggesting that people contribute to biodiversity improvement by species introduction and habitat diversification [4]. On a smaller spatial scale, however, such as urbanization it alters the land use/cover and affects the natural habitat. Densely populated areas are characterized by land fragmentation, isolation of habitats by roads and pollution, and intensively managed agricultural lands [5].

Remotely-sensed data can both be used to evaluate human impact on the biophysical environment and environmental impacts on human economic activity. Understanding human population distribution and density at different spatial levels and landscapes is essential for the formulation of appropriate policies for the sustainable use of natural resources in developing countries. The distribution of human population is one of the key datasets required for improving the understanding of human impacts on land and water resources [6]. Human population in general and population density in particular are often used as proxy measures for land use/cover changes and other spatial changes. Moreover, investigating the linkage between land use/cover and population also helps in examining the risk of natural resource degradation. Therefore, in this paper an attempt is made to reveal the multifaceted relationship between population and land use/cover and to estimate population using remotely-sensed data.

In order to estimate population density a single land use variable (residential or the residential pixel) is normally employed in an ordinary least squares (OLS) regression model to relate population to land use/cover data. It is also possible to relate population to a greater variety of land use/cover types in the form of a multivariate regression model, considering that the population may be found in more than one type of land use/cover [7], [8]. However, there is a spatial non-stationarity issue, which refers to the situation that a relationship tends to vary over space [9]. For population estimation, spatial non-stationarity cannot be addressed using global OLS models. As a result, many researchers have attempted to estimate the population by adopting a regional (local) regression approach, and a pixel-based population estimation approach [7], [10], [11], [12]. Currently, there is an increased interest to estimate population using a geographically weighted regression (GWR) model. The GWR technique has been designed specifically to take care of the above mentioned spatial non-stationarity problem. GWR easily estimates complicated spatial patterns, and is able to capture effects at a local spatial level [9], [13], [14]. However, for rural Africa the spatial non-stationarity relationship between population and land use/cover is rarely investigated. Therefore, this study explores the applicability of a geographically-weighted regression (GWR) model for estimating population densities in rural Africa using land use/cover data that have been derived from remote-sensing while accounting for spatial non-stationarity. Moreover, this study explores the spatial patterns of the GWR model that would help clarify the relationship between population and land use/cover that might not have been evident with the global OLS model. This approach is tested for the case of the Lake Naivasha basin, Kenya.

2. DATA AND METHODS

2.1. Study area

The lake Naivasha basin is located in the central southwest part of Kenya, approximately 80 km northwest of Nairobi, the nation's capital (fig. 1).

It is located in the Kenyan rift valley and is found between latitudes 00° 10' to 00° 55' S and longitudes 36° 09' and 36° 40' E. It covers an area of 3400 km² with a climate that is predominantly semi-arid [15], [16]. The catchment is home for a diversity of flora and fauna, wildlife and bird's habitat that contribute to the area's attractiveness as a tourist destination. Lake Naivasha is registered as an international Ramsar site for wise use of the wetlands through local and national actions and international cooperation to achieve sustainable development in 1995 [17]. Lake Naivasha is a highly significant freshwater resource in an otherwise water deficit area. Apart from the invaluable freshwater, it also supports large and vitally important economic activities including horticulture and geothermal power generation. The upper parts of the basin are mainly used for wheat production and livestock farming. The area to the west and east of the lake are occupied mainly by large-scale farms that produce vegetables and pyrethrum and by maize growing small-scale agricultural farms [16]. Due to land use transformation since Kenya's independence, much of the upper catchment areas of the basin were settled by indigenous Kenyans (fig. 1).

The transformation has continued over the years as large farm areas are sold to land-buying companies, which later subdivided the land into small holdings. In the upper parts of the basin households own up to around 4.04 hectares of land [18]. Agriculture plays a key role in the Kenya's economy. Coffee, tea and horticulture (i.e. flowers, fruits, and vegetables) are the principal exports [19]. The horticultural farms have appeared around Lake Naivasha in the past 20 years depending heavily on the availability of freshwater resources. The area is a major contributor to Kenya's gross domestic product (GDP), for employment opportunities and socioeconomic development of the country as a whole. The growth of large-scale commercial activities in the form of a booming flower industry along with the existing small-scale farms in the Lake Naivasha basin have implications for the demand for resources from the Lake Naivasha basin ecosystem [15].

The population pressure, the intensification of land use in the catchment and the growth of informal settlements in Naivasha town due to lucrative economic activities are major challenges of the Lake Naivasha basin socio-ecological system. As a consequence, more and more people from different parts of the country are attracted to the basin, seeking employment opportunities. Human settlements in the Lake Naivasha basin are concentrated around the main towns, emerging in new rural centres and in farm areas. The basin has experienced significant population growth over the past years and has been estimated at about 568,500 people in 1999 [20]. In 2009, the basin population has been estimated to have increased by approximately 13% reaching values of 659,300 people [21]. A total of 62 sub-locations (the smallest administrative unit) that comprise the census population data of the basin are located within or partly within the catchment area.

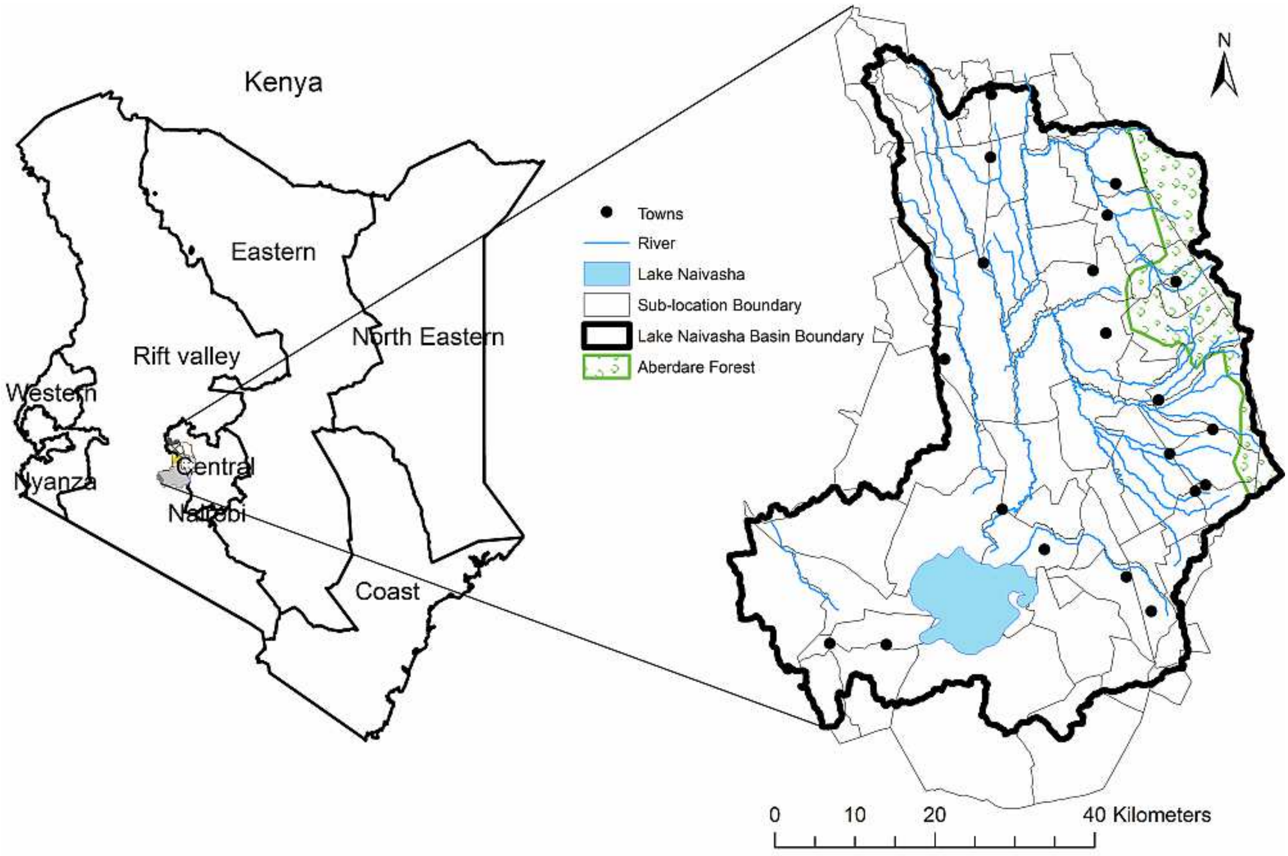

Fig. 1. Lake Naivasha basin, Kenya.

2.2. Data sources

2.2.1. Land use/cover data

A land use/cover map was derived using the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) satellite image[1] (15 m spatial resolution) acquired on 2 February 2008. A geo-referenced and ortho-rectified high resolution (20 cm) aerial photograph of the study area taken between the 3rd and 6th July 2000 by Ramani Geo-systems in the Eastern Africa region was used as a ground truth for land use/cover classification and for visual evaluation of the classification. The aerial photographs covered 300km² ground areas around Lake Naivasha. In addition, an existing data set from Multipurpose Afri-cover Database for the Environmental Resources (MADE) published in 2002 by the Food and Agricultural Organization (FAO) was also used and provided ancillary data to support the classification. A sample of 221 ground truth points were collected from field survey conducted in September 2010 and 56 points were acquired from the aerial photo. An additional 106 survey points were taken from a sample set collected by Were [22] in September 2008 for the same area. Thus, a total of 383 sample land use/cover ground truth points have been collected and used to classify and produce a land use/cover map of the study area for 2009.

The classification of the ASTER satellite image used in this study was implemented using eCognition™ object-based image analysis software. Land use/cover in the Lake Naivasha basin is highly fragmented. As a consequence per-pixel spectral-based methods may not effectively solve the high spectral variation problem within similar land use/cover types. Therefore, an object-oriented classification method was chosen to reduce this problem [23]. Object-oriented image classification and analysis is not focused on single-pixel values, but on group of pixels. Such groups of pixels are called objects. The scale parameter was determined using a method proposed by Dragut et al. [24]. Objects are created in the course of a segmentation process, followed by classification [25]. Image objects can be identified on the basis of patches of spectrally similar pixels referred to as segments [26].

Training data was uploaded as a thematic layer in order to generate sample object for image classification. The standard nearest-neighbour algorithm was used. This algorithm automatically generates multi-dimensional membership functions based on the sample objects. In this process 90% of the data collected from the field survey and the aerial photo were used for image classification. The remaining 10 % of the data was used to create a Training

[1] The ASTER satellite image was obtained from the online Data Pool at the NASA Land Processes Distributed Active Archive Center (LP DAAC), USGS/Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota (http://lpdaac.usgs.gov/get_data).

and Test Area (TTA) mask for assessing the quality of the extracted land use/cover map. The overall accuracy of the classification was 87%, and the Kappa statistic was 83%, indicating a strong agreement or accuracy between classified map and the ground reference information [25], [27], [28]. The overall accuracy result was acceptable and meets a minimum level (85%) of interpretation accuracy needed for the identification of land use/cover categories from remotely-sensed data as recommended by Anderson et al. [29]. The land use/cover map for the Lake Naivasha basin was produced with eight key classes: cropland, forest cover, grassland, woodland, shrub land, built-up area (settlement and urban areas), horticulture (i.e. areas identified as greenhouse for flower and horticulture farms) and water-body. The land use/cover classification result for the whole study area shows that 29.5% of the study area is cropland; 22% is grassland; 24% is forest; 9.2% is woodland; 10.3% is shrub-land; 0.4% is horticulture; 0.7% is built up areas and 3.9% of the study area are covered by water-bodies (fig. 2).

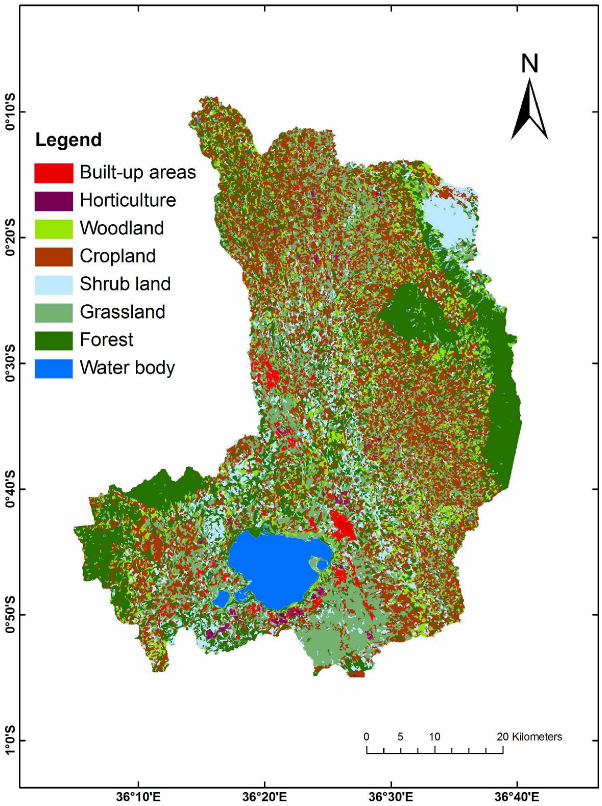

Fig. 2. Land use/cover map of Lake Naivasha basin, 2009.

2.2.2. *Population census data and map*

The 2009 Kenya population and housing census data and the census map (paper copy) were obtained from the Kenya National Bureau of Statistics [21]. The census map was scanned and geo-referenced and co-registered to match the projection of the produced land use/cover map. The 2009 Kenya census population is reported at different administrative units or levels: i.e. national, province, district and sub-location level. The Lake Naivasha basin population census comprises data from 62 sub-locations and Aberdare forest sub-locations in the South and North Nyandarua area are enumerated as sub-locations with no population (fig. 1). These two sub-locations are protected forest areas. The major forest covered areas of the basin are found in these two sub-locations; as a consequence, forest cover is not included in our regression analysis. In addition, the water-body was not included in the analysis.

2.3. Methods

The study employed a global OLS model and a GWR model to estimate population using land use/cover information and to investigate the relationship between land use/cover and population. The global OLS regression model assumes that a stationary relationship between population and land use/cover exists. In spatial processes however, the relationship measurements vary over space, which implies spatial non-stationarity. To address the spatial non-stationarity issue, a GWR model is applied to extend the global OLS model.

2.3.1. *Global ordinary least square (OLS) regression model*

The census population map at sub-location level was overlaid on the land use/cover map in a geographical information system (GIS) following the approach by Yuan et al [12]. An overlay operation with the population map as the destination layer was attached to the land use/cover map; the result was a new map and a data matrix that inherits the attributes from these two maps, which enable us to perform spatial and regression analysis. Let, the data matrix has *m* rows and n+1 columns, where *m* is the number of sub-locations and *n* is the number of land use/cover classes occurring in the sub-locations. The first column of the data matrix is the population count at sub-location level. Then, a global OLS regression model to evaluate the relationship between population and land use /cover type can be written as:

$$Y_i = \beta_0 + \sum_{j=1}^{n} \beta_j X_{ij} + \varepsilon_i \qquad \text{.................................1}$$

Where Y_i is the total population count in the i sub-location for $i=1,2,\ldots,m$, m is the number of sub-locations in the study area, β_o and β_j are the coefficients to be estimated or β_j is the average population density estimate for the j^{th} land use/cover type, X_{ij} is the total area for the j^{th} land use/cover type within i^{th} sub-location in hectare and n is the number of land use/cover classes, and ε_i the random error in population estimate of case *i*.

The random errors represent the variation in population unexplained by the land use/cover types, and it is assumed to be independently and identically normally distributed with zero mean and constant variance δ^2.

2.3.2. Geographically weighted regression model (GWR)

Geographically weighted regression (GWR) is a simple but powerful method for exploring non-stationary spatial relationships. In GWR, all coefficients vary over space, and the parameter estimates are made using an approach in which the contribution of a sample to the analysis is weighted based on its spatial proximity to the specific location under consideration. Data from observations close to the location under consideration are weighted more than data from observations far away.

GWR is a technique that expands standard regression for use with spatial data, and consequently, parameter estimates are more local rather than global parameters.

The theoretical backgrounds and techniques for GWR model have been intensively discussed [9,13,14,30,31]. Suppose we have a standard set of observations X_{ij} for $i=1,2,\ldots,n$ cases and $j=1,2,\ldots,k$ explanatory variables, and a set of dependent variables Y_i for each case. This is a standard data set for a global regression model. Now suppose that in addition to this we have a set of location coordinates (u_i, v_i) for each case. The underlying standard model for GWR is given as:

$$Y_i = \beta_0(u_i, v_i) + \sum_{j=1}^{k} \beta_j(u_i, v_i) X_{ij} + \varepsilon_i \ldots\ldots\ldots 2$$

Where $\beta_0(u, v),\ldots, \beta_k(u, v)$ are $k+1$ continuous functions of the location (u, v) in the geographical study area, and ε_is are random error terms. In the basic GWR model we assume that the error terms are independently normally distributed with zero mean and common variance δ^2 [9].

Accordingly, for this study the GWR model extends the OLS model in Equation 1 by allowing the parameters to be estimated by a weighted least squares procedure. By using a weighting system dependent on the location in geographic space, it allows for local parameter estimates rather than global ones [8]. The GWR model for this study can be written as:

$$Y_i = \beta_{i0} + \sum_{j=1}^{n} \beta_{ij} X_{ij} + \varepsilon_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3$$

Where Y_i is the total population count in *i* sub-location for $i=1,2,\ldots,m$, β_{ij} is the value of the j^{th} parameter estimates (coefficients) to be estimated at sub-location *i* or the average population density estimates for the j^{th} land use/cover type with respect to sub-location *i*, X_{ij} is the total area for the j^{th} land use/cover type within the i^{th} sub-location in hectare and *n* is the number of land use/cover classes and ε_i is the random error in the population estimate of case *i*. As a result, instead of being fixed in the global OLS model, the coefficients β_j are now varying with respect to the sub-location *i*. A geographically weighted Guassian regression has been applied at a sub-location level. To specify the location of each sub-location *i* the latitude and longitude of its centroid are included in the model. The estimated value of the parameter estimates (coefficients) are a function of the bandwidth of the spatial-kernel used, i.e. the radius or the number of observations around each point included in weighting matrix [32]. Two types of spatial kernels are typically used to limit the number of data points considered for each parameter estimates, the fixed spatial kernel and adaptive spatial kernel. In fixed-kernel the bandwidth at each regression point is constant across the study area while an adaptive spatial kernel permits the use of a variable bandwidth [9]. An adaptive kernel produces changing bandwidths that act to ensure that the same number of non-zero weights is used for each regression points in the analysis. This may be more reasonable in addressing spatial non-stationarity [11], [33].

An adaptive spatial kernel was used in this study due to the fact that census sub-locations are varied considerably in terms area and population data. Selection of the weighting function and optimal bandwidth are accomplished by minimizing the corrected Akaike Information Criterion (AICc), which indicates how close a regression model approximates reality, accommodating for differences in the number of degree of freedom in the model compared [9], [34]. Thus, with this parameterization, the bandwidth is the same for all the covariates and the result is the best fitted model. Analysis was performed using the GWR (Version 3.0.18. 2003) software package and mapped using ArcGIS™.

3. RESULTS

3.1. Global OLS regression model of population - land use/cover

The global OLS model was run with intercept (constant term) for proper comparisons of the results with the GWR model. In view of the fact that there will be no population if there is no land use/cover class suitable for persons to live, the global regression model was also fitted without constant (intercept) term. These two models were run using the census population of 62 sub-locations of the study area. As the first attempt the five land use/cover class OLS regression model was estimated with constant term (intercept), using Equation 1 in STATA® Data Analysis and Statistical Software. The following model was obtained:

$$Y_i = 6107.76 + 69.11(built-up) + 1.17(cropland) + 1.80(grassland) - 4.92(shrubland) + 2.42(woodland) \quad (4)$$

Table 1. Summary output for global OLS model regression with constant term (intercept).

Dependent Variable	Population	Number of Observations	62
Mean dependent variable	10,564.1	Number of Variables	6
S.D. dependent variable	8,905.28	Degrees of Freedom	56
R-squared	0.69	F-statistic (6, 56)	25.96
Adjusted R-squared	0.67		
S.E. of regression	5,144.30		

Variable	Coefficient	Std. Error	t-Statistic	Probability	[95% Conf. Interval]
Constant	6107.766	1401.83	4.36	0.000	[3299.562 8915.971]
Built-up area	69.117	18.07	3.82	0.000	[32.91307 105.3213]
Cropland	1.175	1.16	1.01	0.319	[-.1637676 3.771043]
Grassland	1.803	0.98	1.84	0.072	[-1.164582 3.515395]
Shrub land	-4.923	2.14	-2.30	0.025	[-9.217355 -0.628825]
Woodland	2.426	4.27	0.57	0.573	[-6.140195 10.99399]

The regression result of the model reveals that the coefficient for built-up area and grassland cover are positively and significantly related to population at 1% ($p<0.01$) and 10% ($p<0.1$) respectively. Only the coefficient for shrub land has a negative value and is significant at 5% ($p<0.05$), suggesting that its impact on the population is not as great as built-up area and grassland cover. Built-up area has the highest positive coefficient, signifying its strong relationship with the population. In this model the coefficients for cropland and woodland use/cover are positive but insignificant. The model adjusted R^2 value is 0.67 (table 1). Secondly, the five land use/cover class global OLS regression model was also estimated without constant term (intercept) and the following model was obtained:

$$Y_i = 71.31(built-up) + 2.11(cropland) + 2.37(grassland) - 6.62(shrubland) + 9.60(woodland) \quad (5)$$

Table 2. Summary output for global OLS model regression without constant term (intercept).

Dependent Variable	Population	Number of Observations	62
Mean dependent variable	10,564.1	Number of Variables	5
S.D. Dependent variable	8,905.28	Degrees of Freedom	57
R-squared	0.82	F-statistic (5, 57)	53.90
Adjusted R-squared	0.81		
S.E. of regression	6,020.07		

Variable	Coefficient	Std.Error	t-Statistic	Probability	[95% Conf. Interval]
Built-up area	71.313	20.62	3.76	0.000	[30.018 112.608]
Cropland	2.116	1.09	1.94	0.058	[-0.070 3.771]
Grassland	2.372	0.95	2.49	0.016	[0.466 4.278]
Shrub land	-6.622	1.75	-3.77	0.000	[-10.136 -3.107]
Woodland	9.601	3.41	2.81	0.007	[2.765 16.437]
Variable	Coefficient	Std.Error	t-Statistic	Probability	[95% Conf. Interval]

For the Equation 5 regression model, the result reveals that the coefficients for cropland, built-up area and grassland are positively and significantly related to population. Similar to the first model, the coefficient for shrub land cover is negative and significant. Built-up area and shrub land cover are highly significant at 1% ($p<0.01$). Grassland and woodland use/cover are significant at 5% ($p< 0.05$) and cropland use/cover is significant at 10% ($p<0.1$). The *t* value in this model indicates that the five land use/cover variables are significantly related to population as compared to the global model with constant term or intercept. And a significant improvement is shown in adjusted R^2 value from 0.67 to 0.81 (table 2). Compared to the global model with constant term, the explanatory variables (land use/cover types) significantly improved when the global model was estimated without a constant term.

3.2. Geographically weighted regression (GWR) model of population - land use/cover

The GWR model output allows to map the distribution of parameter estimates (coefficients) of the explanatory variables and to test their significance of spatial variability [9].

A GWR model was employed to estimate the coefficients in Equation 3 and to explore the spatial relationships between population and land use/cover in the Lake Naivasha basin. In addition, comparison was made to test whether the local model had a significant improvement over the global model (table 3).

Table 3. Output summary of the local GWR model.

		Global OLS Model	Local GWR Model
Residual sum of squares		1506843448.95	1221136653.90
Effective number of parameters		6.00	9.60
Sigma		5187.28	4827.44
Akaike Information Criterion		1246.40	1243.31
Coefficient of Determination		0.69	0.76
Adjusted r-square		0.64	0.71
ANOVA			
Source	OLS Residuals	GWR Improvement	GWR Residuals
SS	1506843449.0	285706784.0	1221136653.9
DF	6.00	3.60	52.40
MS		79359739.7891	23304200.3003
F	3.405		
Tests of Spatial variability of Parameters			
Parameter	**P-value**		
Intercept	0.72000 n/s		
Built-up Area	0.32000 n/s		
Cropland	0.05000*		
Grassland	0.04000*		
Shrub land	0.11000 n/s		
Woodland	0.21000 n/s		* = significant at 5% level

GWR models, an F-test is used to assess whether spatial variation exists in the relationship under study [32], specifically testing whether the GWR model offers an improvement over, and describes the relationship significantly better than the global OLS model. This was addressed through analysis of variance (ANOVA) test. The F-value is 3.40 suggesting that the GWR model has a significant improvement over the global model in determining the relationship between population and land use/cover.

The corrected Akaike Information Criterion (AICc) of the local GWR model (1243.62) is less than the one of the global OLS model (1246.4) indicating that GWR model performs better than the global OLS model and provides a better fit for observed data (table 3). This is also confirmed by the improvement in adjusted R^2 values.

The significance of spatial variability for each independent variable in local parameter estimates (coefficients) is tested using a Monte Carlo significance test [9]. The result indicates that grassland cover and cropland use/cover are significant at 5% while the other three land use/cover classes are not significant (table 3). In terms of the goodness-of-fit, the GWR model R^2 varies spatially over the entire study area from 0.43 to 0.87. R^2 values and its distribution indicated that around 83% the local estimates is greater than 0.70 as illustrated in Figure 3. Only 17% of the local model has a R^2 value less than 0.69. The population-land use/cover model used with global OLS techniques to produce coefficients that are applied to the whole study area (i.e. for all sub-locations) to estimate population.

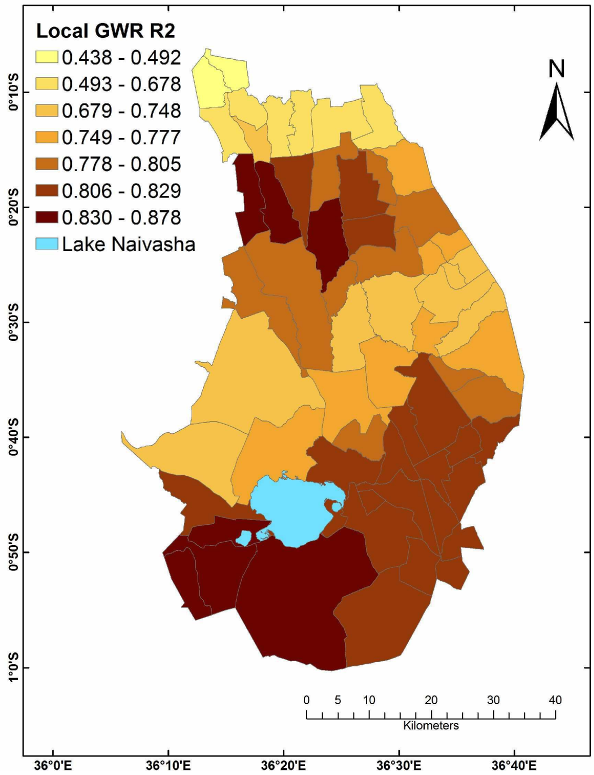

Fig. 3. Local GWR model R^2 values.

However, the population-land use/cover model with a GWR technique can deal with spatial varying relationships and the model produces coefficients for each sub-location in the study area to estimate population. A sample result on estimated population, errors and relative errors for 31 sub-locations for the global OLS model and for the GWR model are presented in Table 4. The accuracy of estimated population based on these two models was evaluated, and it was found that the GWR model produced a more accurate population estimates with a lower root mean square error (RMSE) of ±4827.44 compared to the global OLS model ±5187.28 (table 3).

The GWR analysis in this study was based on a relatively small sample size due to the basin covers a limited number of sub-locations, however, as recommended by Pez, A. et al. concerning a smaller sample size [35], caution has been taken in the analysis to minimize the spurious correlation of the coefficients.

4. DISCUSSION

The two global OLS models indicated that a significant relationship exists between land use/cover types and population in this region assuming stationary relationships. However, the stationary relationship may not be representative for local situations because the land use/cover map extracted from satellite image cannot be 100% accurate and there is spatial variability of classification errors. In addition, population data at the census tract (sub-location) level is also subject to spatial non-stationarity due to the modifiable aerial unit problem (MAUP), which will affect population density in each census tract [8].

Thus, the population-land use/cover model should be spatially non-stationary, which OLS model cannot address. This can easily be explained by the fact that the global model essentially ignores any potential variations in space. Similar to the findings in Langford and Lo that examined the benefits of local model rather than a global regression model in population estimations and for spatial non-stationarity relationship between population and land use/cover [7] [8], this study also justifies the employment of GWR model in rural Africa to have a better estimate of population and to handle the spatial non-stationarity which is not addressed by the global OLS model.

Table 4. Estimated population, errors, and relative errors for selected 31 sub-locations.

List of Sub-Locations	Census population[2]	GWR Model (local model)			Global OLS Model		
		Estimated population	Error	RE%	Estimated population	Error	RE%
MIKARO	3,292	5,100.43	-1,808.43	-0.549	6,897.64	-3,605.64	-0.523
MATINDRI	3,458	6,586.70	-3,128.70	-0.905	5,921.55	-2,463.55	-0.416
RORONI	3,893	4,909.08	-1,016.08	-0.261	5,409.18	-1,516.18	-0.280
KIAMBOGO	4,558	5,785.66	-1,227.66	-0.269	1,874.93	2,683.07	1.431
MAKUMBI	4,679	5,507.86	-828.86	-0.177	2,542.22	2,136.78	0.841
MIKEU	4,807	5,419.70	-612.70	-0.127	2,575.47	2,231.53	0.866
LERESHWA	4,850	6,375.36	-1,525.36	-0.315	3,169.93	1,680.07	0.530
GATAMAIYU	4,914	6,499.09	-1,585.09	-0.323	8,265.37	-3,351.37	-0.405
MALEWA	5,842	7,531.17	-1,689.17	-0.289	9,011.82	-3,169.82	-0.352
GATONDO	6,134	8,406.65	-2,272.65	-0.371	10,342.80	-4,208.80	-0.407
MURUAKI	6,174	8,944.72	-2,770.72	-0.449	5,700.51	473.49	0.083
GETA	6,215	4,809.82	1,405.18	0.226	2,845.41	3,369.59	1.184
KOINANGE	6,323	7,089.60	-766.60	-0.121	4,512.17	1,810.83	0.401
MUNUNGA	6,557	8,448.48	-1,891.48	-0.288	5,964.15	592.85	0.099
KANDUTURA	6,692	10,396.39	-3,704.39	-0.554	12,306.12	-5,614.12	-0.456
KIRIKO	7,481	5,930.53	1,550.47	0.207	5,412.95	2,068.05	0.382
NDEMI	8,176	6,843.21	1,332.79	0.163	13,338.02	-5,162.02	-0.387
KARATI	8,302	10,124.57	-1,822.57	-0.220	6,352.38	1,949.62	0.307
KINJA	8,514	8,253.92	260.08	0.031	5,971.02	2,542.98	0.426
TARAMBETE	8,699	5,364.16	3,334.84	0.383	8,908.91	-209.91	-0.024
MUKUNGI	8,919	9,666.01	-747.01	-0.084	7,155.66	1,763.34	0.246
KINAMBA	9,135	6,646.10	2,488.90	0.272	2,524.40	6,610.60	2.619
MUNYEKI	11,093	12,966.51	-1,873.51	-0.169	12,865.08	-1,772.08	-0.138
MHARATI	12,299	10,604.96	1,694.04	0.138	10,336.42	1,962.58	0.190
NAANDARASI	12,341	8,355.10	3,985.90	0.323	6,849.57	5,491.43	0.802
GITHIORO	12,556	14,838.32	-2,282.32	-0.182	15,972.87	-3,416.87	-0.214
WANJOHI	13,846	10,564.12	3,281.88	0.237	10,691.12	3,154.88	0.295
MWANGO	14,418	12,805.61	1,612.39	0.112	10,808.29	3,609.71	0.334
MURUNGARU	14,709	13,689.76	1,019.24	0.069	14,820.32	-111.32	-0.008
LAKEVIEW	20,082	10,138.92	9,943.08	0.495	2,174.96	17,907.04	8.233
KAHURU	20,803	19,678.88	1,124.12	0.054	22,969.07	-2,166.07	-0.094

[2] The census population is the population count reported at sub-location level in 2009 Kenya population and housing census.

The GWR model produces local parameter estimates (coefficients) for the independent variables by location. The spatial distribution of the coefficients reveals that the relationships between land use/cover variables and population vary not only in magnitude, but also in direction. These parameter estimates show the strength of the relationship of the explanatory variables (land use/cover types) to population by location. The local parameter estimates for built-up area was uniform and shows a positive relationship throughout the study area, with a high positive value (strong relationship) in the northern part and trending down to low positive values (weaker relationship) in the south-western part of the Lake Naivasha basin (fig. 4A). The result pointed to a strong positive relationship between population and built-up area in the northern parts of the basin, which is characterized by the existence of many small towns and new settlement areas.

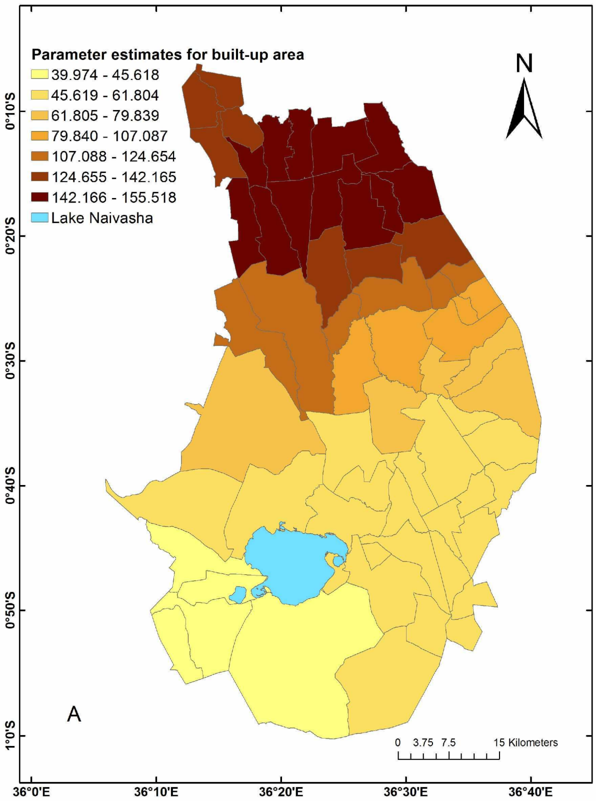

Fig. 4A. Local parameter estimates (coefficients) for Built-up area.

The economy of the basin is anchored in the agricultural sector and experienced a significant population growth over the past 30 years [20,21]. Agricultural activity in the basin has expanded considerably in terms of both the smallholder farmers, which are mainly subsistence farms in the upper catchment with high value commercial horticulture around the Lake [15], [20]. The global OLS model result shows that population is positively related with cropland use/cover while the GWR model estimates indicate that such relationship did not hold in many spatial units of the study area. In the GWR model, there is a high spatial non-stationarity for cropland use/cover and the local parameter estimates change sign over space and highly significant at 5% level in spatial non-stationarity test (table 3).

In the northern middle part of the basin, a strong positive relationship between population and cropland use/cover has been found. This strong positive population-cropland use/cover relationship result is consistent with the study by Becht et al. that indicated the presence and a continuous expansion of smallholder farmers and agricultural farming practices in the upper catchment areas of the basin [15]. However, around the lake and in south-western parts of the basin, the population-cropland use/cover shows a negative relationship due to small-scale agricultural activities are very modest in these areas. Besides, large-scale agricultural farm practices are common around Lake Naivasha. Moreover, the area south of Lake Naivasha is semi-arid which might hinder the expansion of smallholder's farms (fig. 4B).

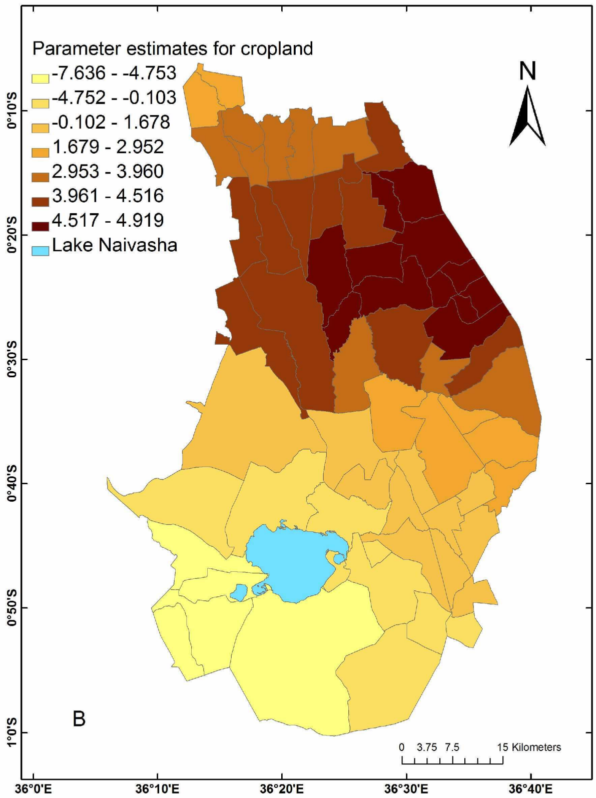

Fig. 4B. Local parameter estimates (coefficients) for Cropland use/cover.

Worldwide, grassland cover is found most commonly in semi-arid zones (28% of the world's grasslands), followed by humid (23%), cold (20%), and arid-zone (19%). Human population is the highest in the dry grassland (arid, semi-arid, and dry sub-humid) areas of Sub-Saharan Africa followed by Asia [36]. Similarly, Lake Naivasha is located in semi-arid zone. Although there is a very rapid rise in large-scale horticultural farms around the lake, these areas are still largely occupied by traditionally pastoral land; the pastoral communities still depend on the lake for watering and grazing their cattle [15], [37]. The major grassland cover is found in the southern, south-western and south-eastern part of the basin with sparsely scattered shrub land cover. Accordingly, the spatial pattern indicates a trend of low values in the north and high values in the south. The local parameter estimates have positive values (i.e., strong relationship) in southern and south-eastern parts of the basin as illustrated in Figure 4C. The northern part of the basin shows negative values, indicating a weak relationship between population and grassland cover. The grassland cover is highly significant at 5% level in spatial non-stationarity test.

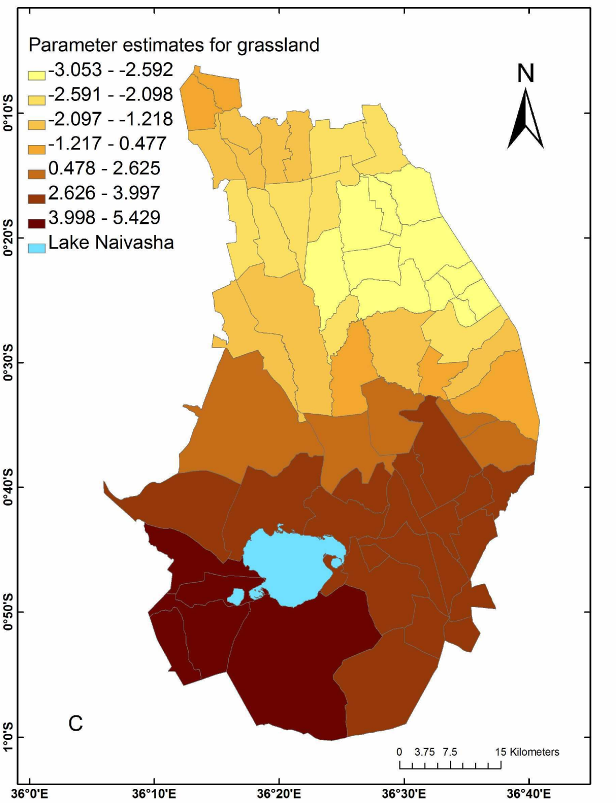

Fig. 4C. Local parameter estimates (coefficients) for Grassland cover.

The population-cropland use/cover and population-grassland cover relationship suggested that the conclusion from the global model may be challenged because of a significant spatial non-stationarity in determining the relationship between population and land use/cover. The result also indicates that considering the spatial variations in the local parameter estimates (coefficients) of the land use/cover variables substantially improved the population estimates (table 4). Therefore, this study indicates that the land use/cover data in rural Africa can provide vital information to estimate population and it can also be used to model the spatial pattern of population density using high resolution satellite imagery.

5. CONCLUSION

This study contributes to improving of population density estimations using remotely-sensed data on land use/cover for the Lake Naivasha basin in Kenya. Remote sensing information on land use/cover was obtained from ASTER satellite data through object-oriented image classification. Two global OLS regression models and a GWR model were applied to explore the relationship between population and land use/cover. The global OLS models failed to deal with spatial non-stationarity. However, the GWR model substantially improved the population estimates by accounting for local variation through the potential spatial non-stationarity. Moreover, the population-land use/cover relationships are better described using a local model than a global model given the spatial variation of the regression coefficients. The grassland and cropland use/cover classes are highly significant and show spatially varying relationship with the population for the Lake Naivasha basin.

This study indicates that applying a GWR model can significantly improve population estimations using remotely-sensed data. Although, the methodology and the model results seem to be very appealing, application of these methods needs to be combined with thorough understanding of other factors such as the presence of infrastructural facilities, employment opportunities, economic and political decisions. These factors can be independent of land use/cover to a large extent. The spatial patterns can also reveal useful information to investigate land use/cover changes, to examine the factors affecting population distribution, and to formulate appropriate management strategies to deal with a population estimate at different spatial scales. Therefore, further research may lead in the application of the method to highlight a potential spatial non-stationarity and the use of GWR as a tool to assist in model development to improve our understanding of spatially-varying relationships of other socioeconomic variables using remotely-sensed data.

6. ACKNOWLEDGMENTS

This research is financed by the Netherlands Organization for Scientific Research, WOTRO Science for Global Development (NWO/WOTRO). We are very grateful to Naivasha sub-region Water Resource Management Authority (WRMA) and other Governmental and Non-Governmental Organizations (NGOs) as stakeholders and partners for this project. We are also indebted to all offices that supported us in providing the available data and information. A preliminary result of this paper was presented on the

conference as: "The spatial non-stationarity of population-land use/cover relationships in Lake Naivasha basin: An application of Geographical Weighted Regression (GWR)". Innovation and Governance Studies (IGS)-Research School for Socio-Economic and Natural Science of the Environment (SENSE) conference on Resilient Societies: Governing risk and vulnerability for water, energy and climate change, 19-21 October 2011, University of Twente, Enschede, the Netherlands.

REFERENCES

[1] **Ningal, T., Hartemink, A. E., Bregt A. K.** (2008), *Land use change and population growth in the Morobe province of Papua New Guinea between 1975 and 2000*. Journal of Environmental Management 87. 117-124.

[2] **Small, C.** (2004), *Global Population Distribution and Urban Land Use in Geophysical Parameter Space*. Earth Interactions 8.

[3] **Smailes, P. J., Argent, N., Griffin, T. L. C.** (2002), *Rural population density: its impact on social and demographic aspects of rural communities*. Journal of Rural Studies 18. 385-404.

[4] **Di Giulio, M., Holderegger, R., Tobias, S.** (2009), *Effects of habitat and landscape fragmentation on humans and biodiversity in densely populated landscapes*. Journal of Environmental Management 90. 2959-2968.

[5] **Niroula, G. S., Thapa, G. B.** (2005), *Impacts and causes of land fragmentation, and lessons learned from land consolidation in South Asia*. Land Use Policy 22. 358-372.

[6] **Tian, Y., Yue, T., Zhu, L., Clinton, N.** (2005), *Modeling population density using land cover data*. Ecological Modelling 189. 72-88.

[7] **Langford, M.** (2006), *Obtaining population estimates in non-census reporting zones: An evaluation of the 3-class dasymetric method*. Computers, Environment and Urban Systems 30. 161-180.

[8] **Lo, C.** (2008), *Population Estimation Using Geographically Weighted Regression*. GIScience and Remote Sensing 45. 131-148.

[9] **Fotheringham, A. S., Brunsdon, C., Charlton, M.** (2002), *Geographically weighted regression: the analysis of spatially varying relationships*. Wiley & Sons, Hoboken, UK.

[10] **Hervey, J. T.** (2002), *Estimating census district populations from satellite imagery: some approaches and limitations*. International Journal of Remote Sensing 23. 2071-2095.

[11] **Yu, D., Wu, C.** (2004), *Understanding Population Segregation from Landsat ETM+ Imagery: A Geographically Weighted Regression Approach*. GIS cience and Remote Sensing. 41. 187-206.

[12] **Yuan, Y., Smith, R. M., Limp, W. F.** (1997), *Remodeling census population with spatial information from LandSat TM imagery*. Computer Environnent and Urban Systems 21. 245-258.

[13] **Fotheringham, A. S., Charlton, M. E. Brunsdon, C.** (1998), *Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis*. Environment and Planning A 30. 1905-1927.

[14] **Huang, Y., Leung, Y.** (2002), *Analysing regional industrialisation in Jiangsu province using geographically weighted regression*. Journal of Geographical Systems 4. 233-249.

[15] **Becht, R., Odada, E. O., Higgins S.** (2005), *Lake Naivasha: experience and lessons learned brief. Managing lakes and basins for sustainable use, a report for lake basin managers and stakeholders,* International Lake Environment Committee Foundation (ILEC), Kusatsu.

[16] **Becht, R., Nyaoro J. R.** (2006), *Influence of groundwater on lake - water management: the Naivasha case*. The 11th world lakes conference, 31 October - 4 November 2005, Nairobi, Kenya., Ministry of Water and Irrigation International Lake Environment Committee (ILEC), .

[17] **Ramsar** (2011), *The Ramsar List of Wetlands of International Importance*, The Ramsar Convention on Wetlands report. http://www.ramsar.org/cda/en/ramsar-documents-list/main/ramsar/1-31-218_4000_0 (accessed 09/09/ 2011).

[18] **Mireri, C.** (2005), *Challenges Facing the Conservation of Lake Naivasha, Kenya*. Topics of Integrated Watershed Management – Proceedings 89, FWU, Vol. 3.

[19] **KNBS and ICFMacro** (2010), *Kenya Demographic and Health Survey 2008-09*. Kenya National Bureau of Statistics (KNBS) and ICF Macro. Calverton, Maryland.

[20] **WWF.** (2011), *Shared risk and opportunity in water resources: Seeking a Sustainable future for Lake Naivasha,* Strategy and Development. WWF Report 2011. PEGASYS.

[21] **KNBS** (2010), *The 2009 Kenya Population and Housing Census: Population Distribution by Administrative Units*. Population Distribution by Age, Sex and Administrative Units, VOLUME IC, Kenya National Bureau of Statistics (KNBS), Kenya.

[22] **Were, K.** (2008), *Monitoring spatio - temporal dynamics of land cover changes in Lake Naivasha drainage basin, Kenya*. Enschede. ITC. 69.

[23] **Lu, D. S., Hetrick, S., Moran, E.** (2010), *Land Cover Classification in a Complex Urban-Rural Landscape with QuickBird Imagery*. Photogrammetric Engineering and Remote Sensing 76. 1159-1168.

[24] **Dragut, L., Tiede, D., Levick S. R.** (2010), *ESP: a tool to estimate scale parameter for multiresolution image segmentation of remotely*

sensed data. International Journal of Geographical Information Science 24. 859-871.

[25] **Lewiński, S.** (2006), *Object-oriented classification of Landsat ETM+ satellite image.* Journal of Water and Land Development 10. 91-106.

[26] **Aplin, P., Smith, G. M.** (2008), *Advances in Object-based Image Classification.* The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. In: Proceeding of Commission VII. ISPRS Congress. Vol. XXXVII. Part B7, Beijing.

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

[28] **Liu, C., Frazier, P., Kumar, L.** (2007), *Comparative assessment of the measures of thematic classification accuracy.* Remote Sensing of Environment 107. 606-616.

[29] **Anderson, J. R., Hardy, E. E., Roach, J. T., Witmer, R. E.** (1976), *Land use and land cover classification system for use with remote sensor data.* In: USGS Professional Paper 964. Washington, D.C.

[30] **Mei, C. L., Wang, N., Zhang, W. X.** (2006), *Testing the importance of the explanatory variables in a mixed geographically weighted regression model.* Environment and Planning A 38. 587-598.

[31] **Fotheringham, A. S., Charlton, M., Brunsdon, C.** (1996), *The geography of parameter space: An investigation of spatial non-stationarity.* International Journal of Geographical Information Systems 10. 605-627.

[32] **Brunsdon, C., Fotheringham, A. S., Charlton, M. E.** (1996), *Geographically weighted regression: A method for exploring spatial nonstationarity.* Geographical Analysis. 28. 281-298.

[33] **Paez, A., Uchida, T., Miyamoto, K.** (2002), *A general framework for estimation and inference of geographically weighted regression models: 1. Location-specific kernel bandwidths and a test for locational heterogeneity.* Environment and Planning A 34. 733-754.

[34] **Burnham, K. P., Anderson, D. R.** (2002), *Model selection and multimodel inference: a practical information - theoretic approach,* Second ed. Springer, New York.

[35] **P ez, A., Farber, S., Wheeler, D.** (2011), *A simulation-based study of geographically weighted regression as a method for investigating spatially varying relationships.* Environment and Planning A 43. 2992-3010.

[36] **Robin White, Siobhan Murray, Rohweder, M.** (2000), *Pilot Analysis of Global Ecosystems: Grassland Ecosystems,* World Resources Institute. Washington D.C.

[37] **Campbell, D., Lusch, D., Smucker, T., Wangui, E.** (2005), *Multiple Methods in the Study of Driving Forces of Land Use and Land Cover Change: A Case Study of SE Kajiado District, Kenya.* Human Ecology 33. 763-794.