*International Journal of Spatial Data Infrastructures Research, 2008, Vol. 3, 73-94*

*Special Issue GI-Days, Münster, 2007*

# Application of Geographically Weighted Regression to Investigate the Impact of Scale on Prediction Uncertainty by Modelling Relationship between Vegetation and Climate*

Pavel Propastin, Martin Kappas and Stefan Erasmi

Department of Geography/Georg-August-University/Göttingen, Germany
ppropas@uni-goettingen.de, mkappas@uni-goettingen.de,
serasmi@uni-goettingen.de

**Abstract**

Scale-dependence of spatial relationship between vegetation and rainfall in Central Sulawesi has been modelled using Normalized Difference Vegetation Index (NDVI) and rainfall data from weather stations. The modelling based on application of two statistical approaches: conventional ordinary least squares (OLS) regression, and geographically weighted regression (GWR). The analysis scales ranged from the entire study region to spatial unities with a size of 750*750 m. The analysis revealed the presence of spatial non-stationarity for the NDVI-precipitation relationship. The results support the assumption that dealing with spatial non-stationarity and scaling down from regional to local modelling significantly improves the model's accuracy and prediction power. The local approach also provides a better solution to the problem of spatially autocorrelated errors in spatial modelling.

**Keywords:** geographically weighted regression, Normalized Difference Vegetation Index, modelling, climate, Sulawesi

## 1. INTRODUCTION

Studies on geographical patterns of vegetation have often been based upon relationships between characteristics of the vegetation activity such as biomass, vegetation cover fraction, leaf area index etc. versus a set of perceived explanatory variables. Commonly, these studies revealed a measure of any indicator for the vegetation activity against a set of environmental determinants, in the main including climatic factors such as precipitation, temperature or growing-degree days, evaporation, soil moisture or others. The Normalized

DOI: 10.2902/1725-0463.2008.03.art6

Difference Vegetation Index (NDVI) derived from multi-spectral satellite data is the most used surrogate of vegetation activity and vegetation characteristics in these studies (e.g. Richard and Poccard, 1998; Yang et al., 1998; Li et al., 2002; Wang et al., 2001; Ji and Petters, 2004). The NDVI is known to be highly correlated to green leaf density, absorbed fraction of photosynthetically active radiation (fAPAR) and above-ground biomass and can be viewed as a general proxy for photosynthetic capacity (Asrar et al., 1984; Sellers et al., 1997; Justice et al., 1985).

One commonly noted feature is that the relationship between vegetation and its spatial predictors appears to vary as a function of geographical region and a number of the underlying environmental factors such as vegetation type, soil type and land use (Wang et al., 2001; Yang et al., 1997; Ji and Peters, 2004). Moreover, the NDVI-climate relationship is also not the same within one land-cover type. There are many cases that show a non-stability of this relationship in space within the same land cover or vegetation type (Fotheringham et al., 1996; Foody, 2003; Foody, 2004; Wang et al., 2005; Propastin and Kappas, 2008). According to these studies, when modelling the spatial vegetation-climate relationship one should take into account that one has to deal with a phenomenon of non-stationarity of this relationship across space. Non-stationarity means that the relationship between variables under study varies from one location to another depending on physical factors of the environment which are spatially autocorrelated. Local regression techniques, such as geographically weighted regression (GWR) help to overcome the problem of non-stationarity and calculate the regression model parameters varying in space (Fotheringham et al., 2002). Because of spatial non-stationarity, the parameters of the model describing the relationship may actually vary greatly in space producing a mosaic that reflects distribution of interaction between the response variable and the predictor factor. This mosaic, however, might demonstrate different patterns at each scale, because different results may be obtained from an analysis by varying its spatial resolution (Openshaw, 1984). Obviously, that the scale-dependent results may be expected with a change in the spatial resolution if a relationship is spatially non-stationary. Spatial variation in the relationship between variables both at and between spatial scales is reported in the recent literature for studies with spatially distributed environmental data. The study by Foody (2003 and 2004), Propastin and Kappas (2008) showed that the predictive power as well as the rank order of explanatory variables in spatial models between remotely sensed data and climatic parameters is a function of scale.

Foody (2004) meant by the term “scale effect” the influence of scale on the outputs of a model (strength of the relationship, parameter values and direction, prediction accuracy, etc.) and suggested that the scale effect is a consequence of the relationship between the variables varying in space. Observations of scale-dependent results can indicate that the explanatory processes and variables operate at different spatial scales. Concerning the spatial distribution of vegetation, the scale effect may be used (1) to analyse variations of microclimate and their effect to vegetation, (2) to determine the minimal size of landscape units reacting to climate factors as a homogeny area, and (3) to find a model with the best prediction power.

In this paper, we analyse scale-dependency of spatial relationships between NDVI and two climatic factors, - rainfall amounts and temperature, - in Central Sulawesi, Indonesia. The aim of the study was: (1) to show spatial variations in the relationship between variables both at and between different scales; (2) to determine the spatial scale at which the NDVI-precipitation modelling achieves the best prediction power and the best prediction accuracy. We tested seven different scales (ranging from the entire study area to local) using two regression techniques - the conventional global OLS regression, and a local regression based on geographically weighted regression (GWR). Certainly, most of the facts discussed in this paper are well known to experts working at the field of geostatistics and have been addressed in geostatistical methods for years. However, local regression models have not yet found a wide acceptance in the international literature on ecology and environmental analysis (for example, see the critical paper addressing local regression methods by Jetz et al., 2005). Therefore, the authors of this paper did not have as an objective to make innovations at the field of geostatistical techniques of data analysis. This paper should be rather considered as a contribution to the current discussion about local and global approaches to spatial data analysis in ecology and environmental science. The results produced in the study should be used in additional modelling the primary production and its dependence on the environmental factors in the Lore-Lindu National Park.

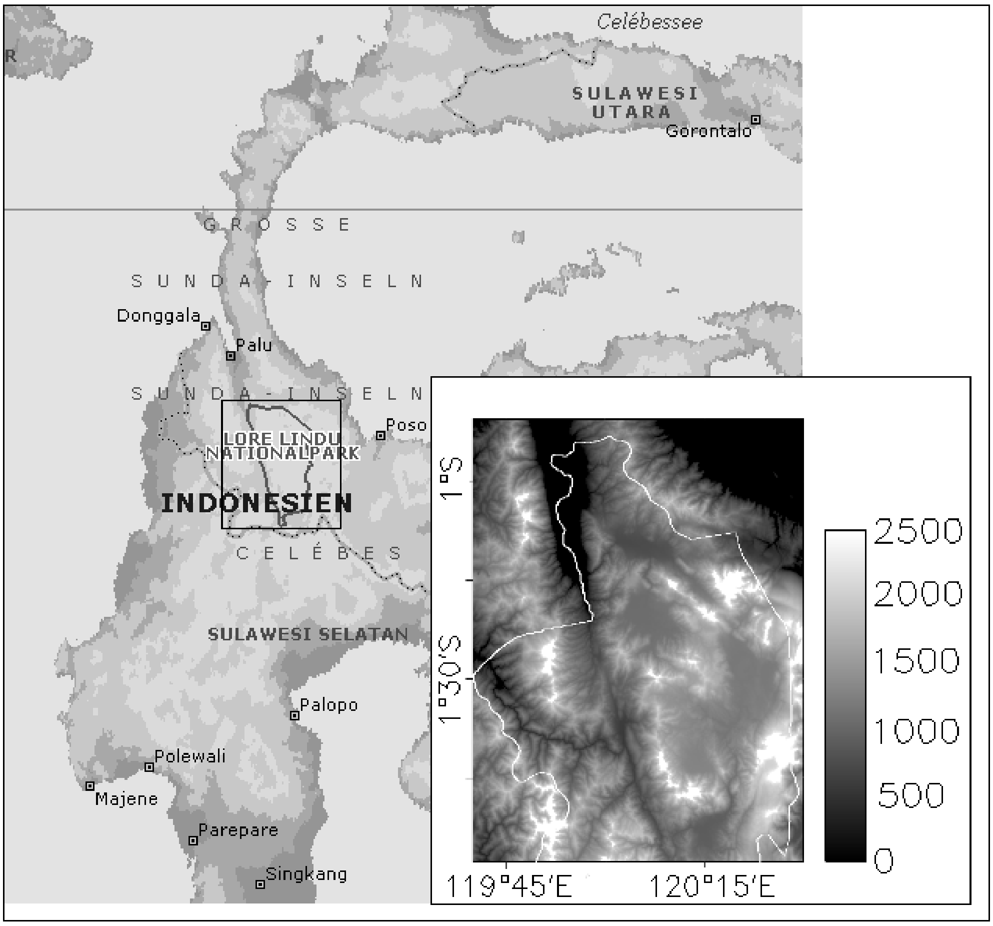
## 2 MATERIALS AND METHODS

### 2.1 Study area

The analysis area is located in Central Sulawesi, Indonesia (Latitude 0°55’-01°54’ South, Longitude 119°40’- 120°29’ East) and comprises the region of the Lore-Lindu National Park together with the bordering areas (Figure 1). The area has a very complicated relief with elevations from zero in the north to more than 2300 m above sea level in the middle part and is cut by four river valleys: the Palolo to the north, Napu to the east, Bada to the south and Kulawi to the west. The highest peaks are Mt. Nokilalaki (2355 m) and Mt. Rorekatimbu (2610 m).

In terms of the climate type, the study area belongs to the belt of equatorial humid climate. The annual rainfall ranges from about 2000 mm in the north to more than 3000 mm in the south. It falls throughout the year and the heaviest period is during the northern monsoon which lasts from November to April. There is no pronounced wet and dry season. The daytime temperature in lowland areas of the region ranges from 26-28°C throughout the year. However, due to the complex terrain and the diverse geomorphological setting the climate is characterised by large spatial variations. For instance, the main valley of the Palu River receives only 600-800 mm precipitation, while mountain slopes east and west of the valley may have up to 2500-3000 mm of annual precipitation. The spatial distribution of mean daily temperature is also depending strongly on elevation and falls in the mountainous areas to 15-16° C.

**Figure 1: Maps present the location of the study area in Sulawesi (left) and its relief (elevation above sea level in m). The white line on the right map shows the border of the Lore-Lindu National Park.**



The natural vegetation is generally classified into two major vegetation types based on altitudinal distribution with lowland rainforest below 1000 m and mountain rainforest above 1000 m. Most areas of the river valleys are completely deforested and used for production of paddy rice; the most common upland cropping systems in the research area are maize and perennial agroforestry systems with cocoa and/or coffee (Whitten et al., 2002).

## 2.2 NDVI dataset

The most recent studies on spatial and temporal relationships between vegetation and climate at global or regional scales have been based on the using of the satellite derived Normalized Difference Vegetation Index (NDVI). The NDVI is established to be highly correlated to green-leaf density, absorbed fraction of photosynthetically active radiation and above-ground biomass and can be viewed as a major surrogate for vegetation activity (Tucker and Sellers, 1986). The vegetation absorbs a great part of incoming radiation in the red portion of the spectrum (R=380-730 nm) and reaches maximum reflectance in the near-infrared channel (NIR=730-1100 nm). The NDVI, defined as ratio (NIR-R)/(NIR+R), represents the absorption of photosynthetic active radiation and hence is a measurement of the photosynthetic capacity of the canopy. Negative NDVI values indicate non-vegetated areas such as snow, ice, and water. Positive NDVI values indicate green, vegetated surfaces, and higher values indicate increase in green vegetation.

This study used NDVI data products with the spatial resolution of 250 m obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS). The dataset covered a time period of four subsequent years from January 2002 to December 2005 and was composed as maximum 16-day values (the entire data set comprised 93 16-day images). Although the use of maximum values significantly reduces noise due to atmospheric effects, particularly, amount of clouds in the dataset (Holben, 1986), the MODIS data over the study region comprised many areas whose NDVI values were contaminated by clouds. The removal of the remained clouds from the 16-day NDVI time series was achieved by using a filtering algorithm based on a weighted least-squares regression approach described by Savitzky and Golay (1964). This algorithm has been successfully used for Spot-VEGETATION data by Chen et al. (2004) and improved for MODIS data by Erasmi et al. (2006). From the filtered MODIS NDVI 16-day data sets, we computed a mean NDVI for the whole period.

## 2.3 Climate dataset

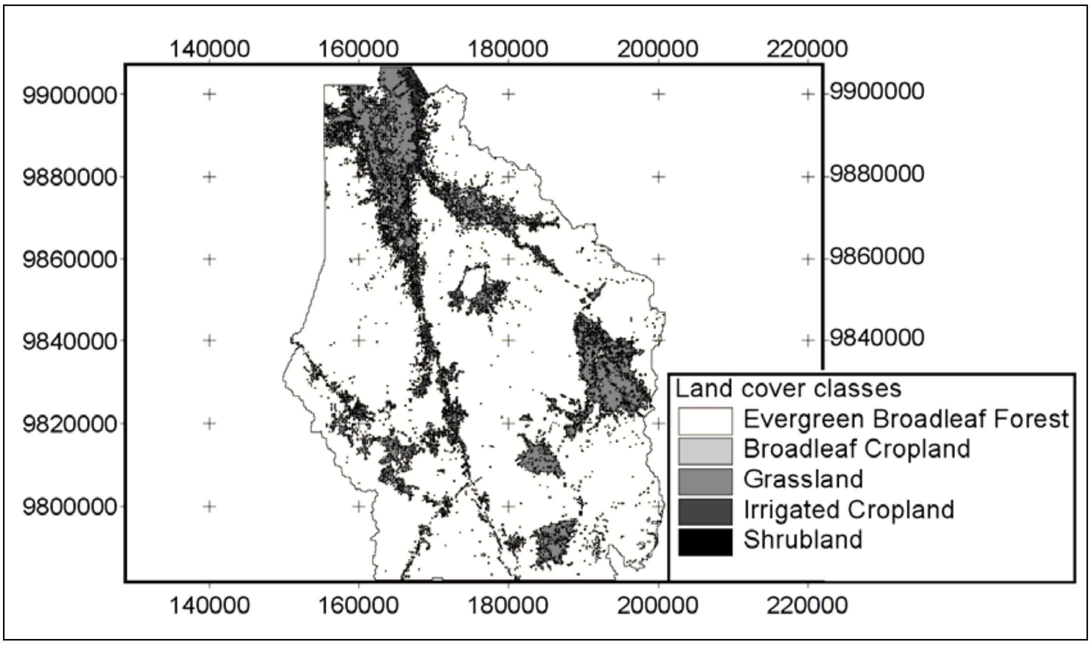
The climate data in the study consist of daily rainfall data and air temperature records collected from a number of automatic climate stations placed throughout the study area. The daily rainfall data were summed to monthly values, while the daily temperature data were averaged to mean monthly values, and gridded maps of precipitation amount and mean temperature for each month from 2002 to

2005 were obtained by interpolating data between the stations using kriging with external drift (Chiles and Delfiner, 1999). Elevation of the climate stations above sea level was used as external explanatory variable. After that, we calculated the average annual precipitation amount and mean annual temperature over the period of 2002-2005.

## 2.4 Land cover

The land cover data in the study area were taken from a digital land-cover map derived from Landsat ETM+ data by Erasmi et al. (2007). The map reveals 12 land cover types in the study area which could be generalized to 5 mean land cover categories used as stratification units in this study: evergreen broadleaf forest, broadleaf cropland (cacao and coffee areas), irrigated cropland (rice areas), grassland and shrubland (Figure 2). Maize was included in the class of broadleaf cropland.

**Figure 2: Land use/land cover in the study area.**



## 2.5 Regression models

Relationships between NDVI as dependent variable and precipitation and temperature as two predictor variables were modelled by using conventional ordinary least squares (OLS) and geographically weighted regression (GWR) analysis. The first one was fitted to the whole study region (global OLS). The second one uses the location information for each observation and allows the model's parameters to vary in space. The GWR was performed with 9 different kernel sizes (from 50*50 pixels to 5*5 pixels). The land area represented by each pixel was 62500 m² (0.625 ha).

As OLS analysis has been well documented in a huge number of textbooks for statistics (for example, Norcliffe, 1981), we just briefly describe the theoretical background for the GWR method. A full description of the geographically weighted regression and its treatments is provided by Fotheringham et al. (2002) or Paez et al. (2002 a, 2002 b).

The simple linear model, usually fitted by ordinary least squares methods (OLS), is:

$$y = a + \beta * x + \varepsilon \qquad (1)$$

where *a* is the intercept of the line on the *y* axis (where *x* = 0), *β* represents the slope coefficient for independent variable *x*, and *ε* is the deviation of the point from the regression line. Fitting the best-fit regression model incorporates the problem to find *a* and *β* so that the total error $\sum \varepsilon_i^2$ is minimized.

Before we continue, we should clarify the terminology that will be used. In this paper we mean by the term "regression parameters" only the coefficients for independent variables: the intercept of the regression line with the x-axis and its slope. By the term "data point" an individual data point with its quantities - value and location – used to fit a regression model is understood. In the case of using raster data, a data point is an individual pixel. The term "regression point" refers to an individual data point which is located in the centre of a mowing window (kernel) used for calibration of a local regression model (see Figure 2).

In the OLS model, the two variables to be related to are *y*, the dependent variable (for this study - NDVI), and *x*, the independent variable (rainfall). The regression model parameters *a* and *β* derived by the above approach are assumed to be stationary over the analysis space (the whole study region or the geographical space occupied by a land-cover type). In other words, applying the conventional global regression model to study the relationships between vegetation distribution and its conditions and environmental parameters, our calculation is based on the assumption, that at each point of the study area this model is absolutely representative and the quantified relationship is constant.

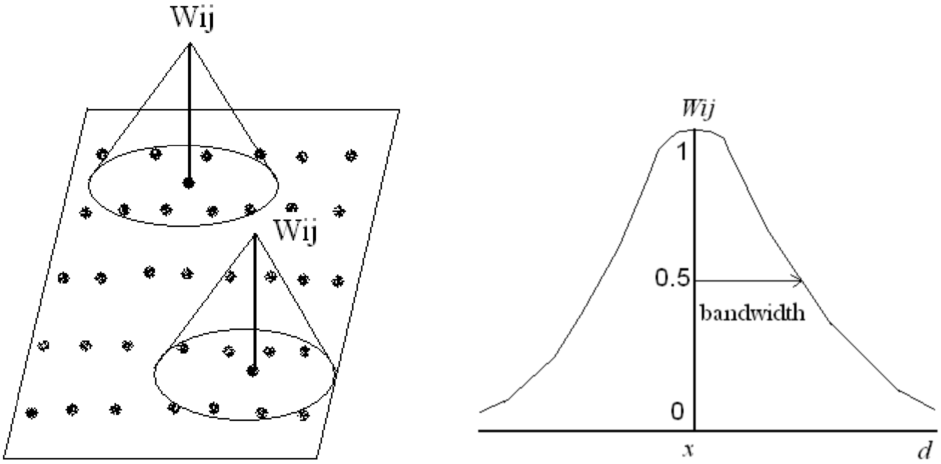
GWR is a local regression technique that allows the model parameters to vary across the space. As our introduction shows (Wang et al., 2001; Yang et al., 1997; Ji and Petters, 2004), in the case of land cover types with different response of vegetation to climate and a diverse orographic environment it is incorrect to hold that the same linear relationship is appropriate in all places. Although the local technique does not allow extrapolation beyond the region in which the model was established, it does allow the parameters to vary locally within the study area and may provide a more appropriate and accurate basis for descriptive and predictive purposes as it has already been shown for NDVI and climate predictors in studies by Foody (2003), Wang et al. (2005), and Propastin and Kappas (2008).

The local estimation of the parameters with GWR is given by the equation (for two independent variables):

$$y = \beta_0(\mu,\nu) + \beta_1(\mu,\nu)x_1 + ..... + \beta_n(\mu,\nu)x_n + \varepsilon \qquad (2)$$

This regression equation orders the regression parameters to be estimated at a location for which the spatial coordinates are provided by the variables $\mu$ and $\nu$. Parameters can also be estimated at locations where there are no data.

**Figure 3. Left: GWR with fixed spatial kernels. In this example, a region is described around a regression point (located in the centre of the mowing window) and all the data points within this region or this window are then used to calibrate the model. Right: A spatial kernel. The variable x represents a regression point, d is the distance between regression point *i* and data point *j*, $w_{ij}$ is the weight of data point *j* at the regression *i*. (the graphs done by the authors).**



The regression parameters may be thought of as a three-dimensional surface over the geographical area rather than a single, fixed real number obtained by the global OLS model. In this model, the regression and its parameters in each point of space is quantified separately and independently from other points. The regression model is calibrated on all data that are positioned within the region described around a regression point and the process is repeated for all regression points (Figure 3). The resulting estimates of the local parameters can then be mapped at the locations of the regression points to view possible non-stationarity in the relationship being examined. The size of the moving window (kernel) is less than the region size and can be varied from one point to another depending on the density of observations at certain area. If the density of observations is greater, the size of the moving window can be diminished. On the contrary, in the sections where the density of observations is low, the mowing window can be enlarged.

The parameters for GWR may be estimated by solving the matrix equation:

$$\hat{\beta}(\mu,\nu) = (X^T W(\mu,\nu) X)^{-1} X^T W(\mu,\nu) y \qquad (3)$$

where $\hat{\beta}$ are intercept and slope parameters in location ($\mu$, $\nu$) and $W(\mu,\nu)$ is a weighting matrix whose diagonal elements represent the geographical weightings of observations around point *i*:

$$W(x_i) = \begin{pmatrix} w_{i1} ...0.....0 \\ 0...w_{i2}.....0 \\ ................ \\ 0....0......w_{in} \end{pmatrix} \qquad (4)$$

where $w_{in}$ is the weight assigned to the observation at location *n*.

Geographically weighted regression works in the way that each data point is weighted by its distance from the regression point. The closer is a data point to the regression point, the more weight it receives. This means that a data point closer to the regression point is more profound in the local regression than are data points located far away. Spatial weighting function can be calculated by different methods. For fixed kernel size, the weight of each point can be calculated by applying Gaussian function:

$$w_{ij} = \exp[-1/2(d_{ij}/b)]^2 \qquad (5)$$

where $d_{ij}$ is the distance between regression point *i* and data point *j*, and *b* is referred to as a bandwidth. Gaussian weighting function is the most used one for application with the data with a regularly distribution in space, because it provides a continuous weighting function up to distance b from the regression point and then zero weights any data point beyond b. Raster data have a very regular distribution and can be analysed using fixed kernel size. In this case, Gaussian weighting function is the most appropriate one (Brunsdon et al., 1996; Fotheringham et al., 2002).

## 2.6 Testing for spatial non-stationarity

Of course, not all relationships exhibit spatial non-stationarity, and the parameters of interest do not vary geographically in all models to be computed and in all geographical regions to be analysed. We may propose that the model parameters should remain stationary in a great number of cases. Fotheringham et al. (2002) described two approaches to test significance of spatial variations in local parameters of a particular data set. The significance of variability in the local estimates can be examined adopting the Monte Carlo test or the Leung test (Brunsdon et al., 1996; Leung et al., 2000). These approaches work in the following simple way: a GWR estimate of the coefficient of interest is taken at each of the n data points and the variance or standard deviation of these estimates is computed. Under the null hypothesis that the model with globally fixed parameters holds, if there are no spatial variations in the parameter, then any permutation of the regression variables against their locations is equally likely and the distribution of the variance should be null. If standard deviation of a local regression estimate is larger then a certain threshold (usually $\pm 1$ standard deviation) assigned by the probability of observing variation in local parameter estimates from a stationary process, than this parameter should exhibit a significant degree of spatial non-stationarity.

## 2.7 Scale issue in the GWR approach and its implications

Estimated parameters in geographically weighted regression depend on the weighting function of the kernel selected. When the bandwidth *b* becomes larger, the model solution will be closer to that of global OLS one. Conversely, as the bandwidth decreases, the parameter estimates will increasingly depend on observations in close proximity to regression point *i* and have increased variance. The problem is therefore how to select an appropriate bandwidth. Obviously, the selection of an appropriate bandwidth for GWR refers to the problem what is the appropriate scale at which to analyse the data. The GWR approach enables a solution of this problem through a number of criteria that can be used for bandwidth selection.

The selection of bandwidth can be determined using the Akaike Information Criterion (AIC) (Hurvich et al., 1998). Minimising the AIC provides a trade-off between goodness-of-fit and degrees of freedom. The AIC is defined for GWR as the following (Fotheringham et al., 2002):

$$AIC = 2n\log_e(\hat{\sigma}) + n\log_e(2\pi) + n\left\{\frac{n + tr(S)}{n - 2 - tr(S)}\right\} \qquad (6)$$

Where *n* is the sample size, $\hat{\sigma}$ is the estimated standard deviation of the error term, and *tr(S)* refers to the trace of the hat matrix which is a function of the bandwidth (Fotheringham et al., 2002). As a general rule, the lower the AIC, the closer is the approximation of the model to the reality. Thus, the best model and the most appropriate scale to analyse the data is the one with the smallest values of AIC. The criterion can also serve for indicating the goodness-of-fit of a particular model and assessment of its appropriation. In this study we computed the value of AIC and used it in a comparison of different models. This criterion will be also used to demonstrate a dependency of the model's goodness-of-fit on the bandwidth.

## 2.8 Uncertainty assessment

Commonly, the results obtained from different models are compared by the amount of variance explained by the corresponding regression model. A general rule is that the higher is $R^2$, the deeper is the understanding of the variables responsible for the variation in the dependent variable. A goodness-of-fit of a regression model increases with the increase of $R^2$. However, using $R^2$ for interpreting the results of GWR does not make much sense, because it will automatically be very high when choosing a small enough band width. Only an $R^2$ adjusted for the degrees of freedom makes sense for inter-comparison of models with different band width. Therefore, the Akaike Information Criterion, AIC, (Equation 6), was used as the major guide to the prediction power of the models. Nonetheless, the values of $R^2$ and the adjusted $R^2$ have been computed in this work and were used for demonstration of the variation in the NDVI-climate relationship between different spatial scales.

Regression residuals contain the very important information about the prediction correctness of a regression model. As the source data demonstrate a strong spatial autocorrelation, a regression modelling with these data is problematic and requires a careful treatment of this phenomenon. To consider the spatial autocorrelation in NDVI-rainfall analysis is of ecological significance can lead to nearby sites in space tending to have more similar values than would be expected by chance. Spatial autocorrelation of the source data makes an application of classical statistical tests like OLS regression for violating the assumption of independently distributed errors problematic. In this case spatial distribution of the regression residuals serve as a significant indicator for the model's uncertainty. An independent distribution of residuals over the analysis space is the sign for a non-problematic regression model. Spatial patterns of regression residuals containing positive autocorrelation indicate that a created model is problematic: the standard errors are underestimated and the correlation coefficient often indicates a significant relationship between variables when in fact there is none (Clifford et al., 1989). In this study, the Moran's *I* coefficient was used as a measure of autocorrelation for the regression residuals. Under the null hypothesis of no spatial autocorrelation, Moran's *I* has an expected value near zero, with positive and negative values indicating positive and negative autocorrelation, respectively. We computed and compared Moran's *I* autocorrelation for residuals from each regression model, the lower the autocorrelation of the residuals, the better is the model.

**Table 1: Summary of the fitting characteristics for the regression models analysed in the study. The best-fitted model is the GWR one with a bandwidth of 1750 meter.**

| Regression model | $R^2$ | Adjusted $R^2$ | RMSE | AIC | Moran's *I* autocorrelation of residuals (distance, pixel/m) |
|---|---|---|---|---|---|
| Global OLS | 0.26 | 0.26 | 0.0667 | -12247.59 | 60 (15000) |
| GWR with the bandwidth, *b*, | | | | | |
| (meter) | 0.51 | 0.50 | 0.0563 | -14231.64 | 34 (8500) |
| 25000 | 0.64 | 0.62 | 0.0473 | -15258.79 | 15 (3250) |
| 10000 | 0.68 | 0.66 | | -16507.03 | 10 (2500) |
| 6250 | 0.76 | 0.74 | 0.0421 | -17262.79 | 5 (1250) |
| 4700 | 0.82 | 0.80 | | -19473.56 | 4 (1000) |
| 2750 | **0.90** | 0.88 | 0.0386 | **-21463.91** | **3 (750)** |
| **1750** | 0.94 | 0.89 | | -21086.51 | 5 (1250) |
| 1500 | 0.97 | 0.90 | 0.0301 | -19033.77 | 8 (2000) |
| 1000 | 0.98 | 0.91 | **0.0210** | -5448.62 | 11(2750) |
| 750 | | | 0.0204 | | |
| | | | 0.0161 | | |
| | | | 0.0119 | | |

## 3. RESULTS AND DISCUSSION

Conventional OLS model fitted to all vegetated pixels of the study area (study area scale) and GWR model with different bandwidth (from 25000 to 750 m) have been adapted to analyse NDVI relationship to the both explanatory climatic variables. Correlation analysis of NDVI with precipitation and temperature revealed that NDVI had significant correlation ($p < 0.05$) with the climatic parameters within all the models above, but the strength of this correlation, prediction power as well as prediction uncertainty of the models showed high variation between the modelling scales. Table 1 shows the derived characteristics for each model. The best model determined from our analysis was the GWR model which had a bandwidth of 7 pixels (1750 m). The worst model was the OLS regression fitted at the scale of the study area.

### 3.1 Study area scale

Table 2 shows the results of the multiple (multivariate) linear OLS model between NDVI, precipitation amounts and temperature comprising all vegetated pixels in the study area. The derived parameters of the model exhibited very high values of the T coefficient, but, they also have relatively high values of standard error. The Monte Carlo test revealed a presence of spatial non-stationarity in the OLS parameter estimates. The spatial variance in the regression parameter estimates was statistically significant at the level of $p < 0.0001$ for each of the parameters.

The results of the global regression model suggest that across the study region NDVI is positively related to precipitation and temperature, but the huge amount of the variance in NDVI remains unexplained. This fact and the presence of non-stationarity in the derived regression parameters means that the model does not adequately represent the real relationship between spatial patterns in vegetation and the climatic factors, and this may drive further work that aims to increase the understanding of the variables responsible for the variation in the dependent variable observed. One of the reasons for the confusion of the global OLS regression to model this relationship may be the wrong spatial scale used for the modelling. The fact that only a very low amount of variance was explained by the OLS model encouraged us to undertake additional investigations which aimed to increase the understanding of the relationship between the variables. The further analysis suggests a link between the strength of this relationship and the physical conditions of the underlying factors, particularly vegetation type and composition of vegetation communities.

**Table 2. Results of the multiple global OLS model between NDVI, precipitation and temperature. The final column indicates the statistical significance of the spatial non-stationarity in the local parameter estimates derived from a Monte Carlo test.**

| Parameter | Estimate | Standard error | T-value | Non-stationarity, *p* < |
|---|---|---|---|---|
| Intercept, $\hat{\beta}_0$ | 0.39772 | 0.116822 | 34.03 | 0.0001 |
| Precipitation, $\hat{\beta}_1$ | 0.00019 | 0.000048 | 41.07 | 0.0001 |
| Temperature, $\hat{\beta}_2$ | 0.00564 | 0.003083 | 18.29 | 0.0001 |

## 3.2 GWR model

By accommodating spatial non-stationarity into the model, the GWR analysis allowed the parameters of the models to vary in space and showed considerably stronger relationships with NDVI than from the corresponding conventional global and stratified regression analysis. It was apparent, that the explanatory power of the models varied between model's scales, with overall estimates of the adjusted $R^2$ varying from 0.50 to 0.91 depending on the used bandwidth. A plot of bandwidth against AIC (Table 1) suggests an optimal value for bandwidth of 1750 meter. Even though supplementary decrease of the bandwidth can give a higher value of $R^2$ (the finest scale studied was 750 meter) and relatively smaller RMSE but it also results in a significant increase of the AIC value. As the bandwidth gets smaller, the degrees of freedom in each local model calibration will decrease and will lead to unstable regression results. Therefore, very small kernel sizes introduced more bias in the model and are characterized by higher values of AIC and underestimated RMSE in comparison to the most appropriate bandwidth because they miss the true scale of spatial variation in the relationship.

The local parameter estimates from the GWR model vary in magnitude and direction, their spatial pattern illustrate the geography of the relationships between NDVI and the climatic factors. Table 3 summarizes the descriptive statistics obtained for the parameter estimates of the GWR model obtained using a bandwidth of 1750 m. Thus, the intercept parameter $\hat{\beta}_0$ has a median of 0.6179 with a range of -2.4320 to 2.7463; the precipitation parameter $\hat{\beta}_1$ has a median value of 0.0001 with a range from -0.0007 to 0.0018. The temperature parameter varied from -0.0822 to 0.0291 with a median of 0.0030. The GWR model exposed the presence of non-stationarity not only between different land-cover categories but also within each of these categories. GWR works in the way that it blows out the arbitrary boundaries between the land-cover categories and represents the NDVI-climate relationship as a continuous geographical process.

However, GWR does not distract the general nature of this relationship, preserving the general differences in the vegetation response to precipitation between individual land-cover categories proved by the stratified OLS model.

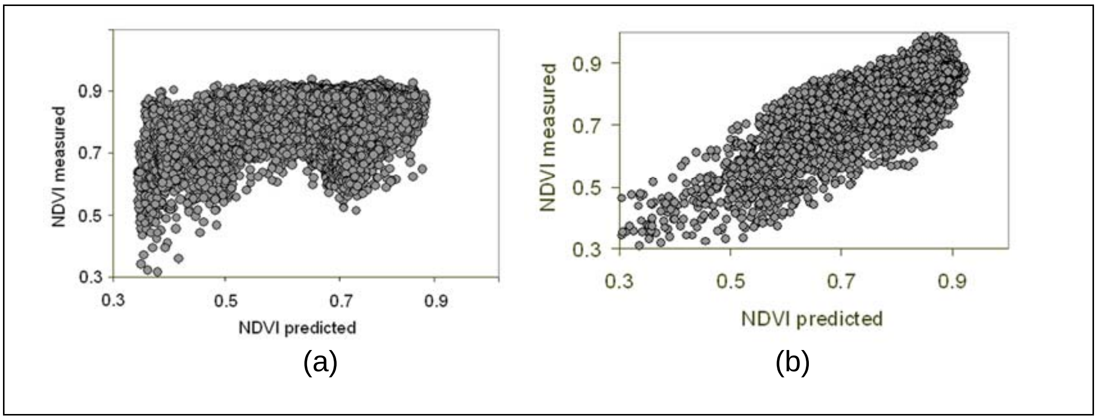
Figure 4 shows the scatter plot between measured NDVI and NDVI predicted using the GWR model with a bandwidth of 1750 m. The results indicate a major degree of spatial variation in the relationship between NDVI and the climatic factors in this region. There are many reasons for this, relating to variations in land cover, vegetation type and composition, terrain features as well as issues connected with the generation and accuracy of the data sets used. The GWR analysis proved that the local variation in this relationship would not be taken into account in an explanation model basing on a conventional, global, OLS regression analysis. Moreover, spatial variations in the amount of unexplained variance in NDVI, which were modelled with the GWR analyses, indicate that the model's prediction power and accuracy are not constant across the study region and vary both between different land cover types and between localities. The global estimates of the regression parameters derived by all the global OLS models used in this study fail to represent the relationship between the analyzed variables at most of the space points and consequently has lesser descriptive and predictive power.

When observing the distribution of the $R^2$ across the study region one can recognize the general pattern which agrees with the land-cover pattern. But in comparison to the stratified model the GWR model also exposes a mosaic of variance in $R^2$ within this general pattern scaling down to the individual locations. It means that the general nature of the relationship appears relatively stable according to the response of different vegetation types to rainfall. Nevertheless, the local variances in this response caused by the variance in underlying physical factors are also included in the model.

**Table 3. Descriptive statistics of the parameter estimates for the GWR model using a bandwidth of 1750 m.**

| Parameter | Minimum | Lwr quantile | Median | Upr quantile | Maximum |
|---|---|---|---|---|---|
| $\hat{\beta}_0$ | -2.4320 | 0.0892 | 0.6179 | 0.8570 | 2.7463 |
| $\hat{\beta}_1$ | -0.0007 | 0.0000 | 0.0001 | 0.0004 | 0.0018 |
| $\hat{\beta}_2$ | -0.0822 | -0.0030 | 0.0030 | 0.0064 | 0.0291 |

**Figure 4: Scatter plot of measured NDVI versus NDVI predicted by (a) the OLS model; and (b) the GWR model with a bandwidth of 1750 m.**

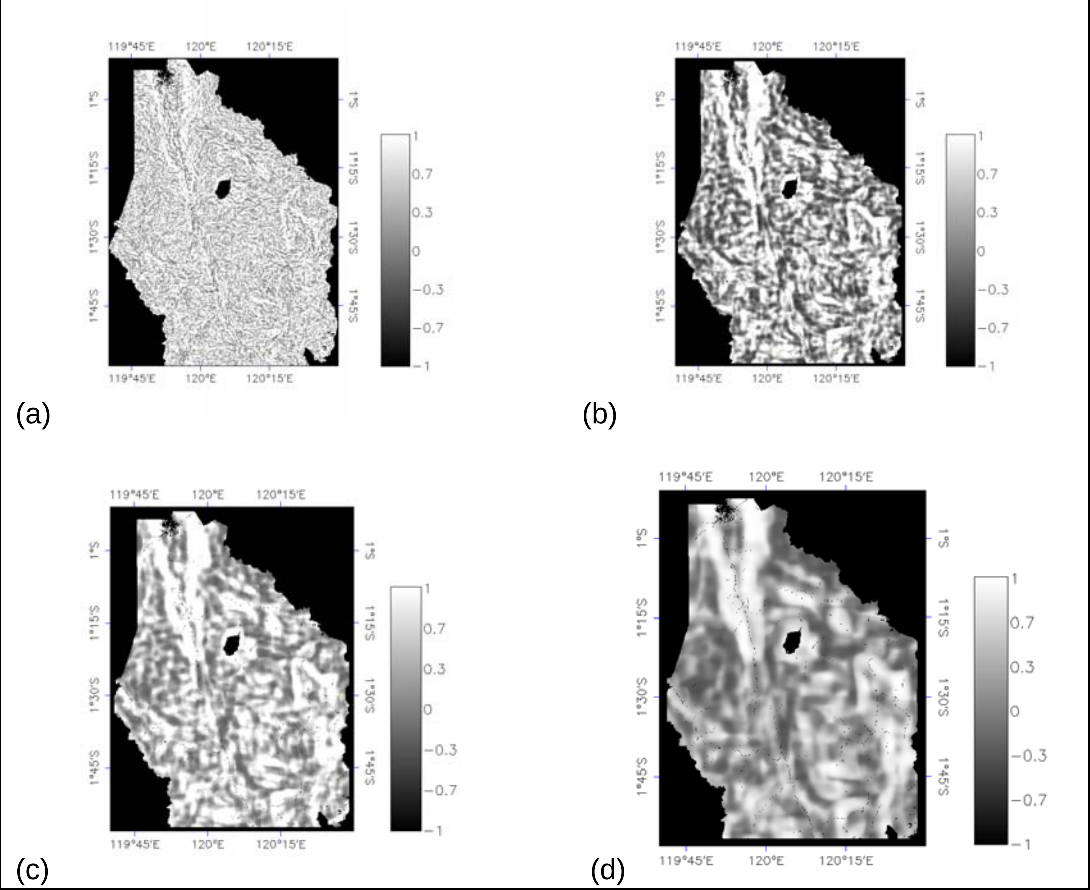


### 3.3 Effect of scale variation on results of the GWR model

In this work the effect of variation in spatial scale of the regression model's prediction was explored by varying the bandwidth used in the GWR model. As the bandwidth declines, the analysis becomes increasingly local, revealing greater geographical detail. With a very small bandwidth (*b* = 750), the relationship between NDVI and precipitation was very strong, the value of the adjusted $R^2 = 0.98$. With an increasing of the bandwidth the spatial patterns in the local estimates of the model parameters became more generalized and the value of the estimated parameters tended towards the global model estimate (Figure 5).

Large windows improve the *t*-value of the model parameters (in sampling terms, increasing the size of the sample) by borrowing large amount of local information but at the expense of introducing bias because information is being borrowed from areas, further away, that may be different. Small windows reduce the risk of bias in the statistics but because little information is being borrowed the precision is not much improved. The effectiveness of local borrowing depends on the local homogeneity of the spatial data which depends on the size of spatial units in relation to the true scale of spatial variation. If adjacent areas are very different in nature then borrowing information locally may introduce bias that distorts the underlying patterns through inappropriate bandwidth. The GWR analysis showed that the most appropriate bandwidth is 1750 meter. This dimension may be considered to reflect the normal size of homogeny landscape units in the study region. The presence of the scale effect in the strength of the NDVI-precipitation relationship and the prediction uncertainty of the model indicates that non-stationarity plays an important role in the ecological modelling and that the geography matters and location should be considered as a variable.

**Figure 5. Spatial variation in the correlation coefficient between NDVI and the climatic predictors at five bandwidths: (a) *b* = 750 m, (b) *b* = 1750 m; (c) *b* = 2750 m; and (e) *b* = 6250 m. Spatial detail increases with a decrease in bandwidth, *b*.**



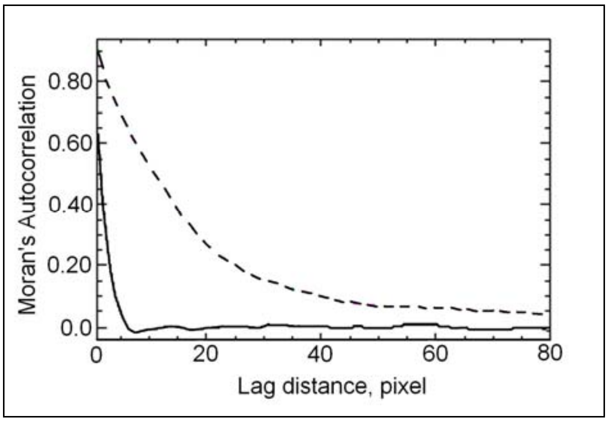
## 3.4 Autocorrelation of regression residuals

For each regression model we calculated the Moran's *I* autocorrelation of the residuals to examine the effect of calibrating the model between NDVI and precipitation at different spatial scales. As it has been proved, the local calibration solves much of the problems of spatially autocorrelated error terms included in the traditional global OLS model (Wang et al., 2005; Fotheringham et al., 2003). We were interested in the comparison of the results from the global and local models. Figure 6 shows the spatial autocorrelograms for the global OLS model residuals and the residuals from the GWR model. As expected, the error terms are most strongly autocorrelated for the global OLS model. The OLS model residuals had significant spatial autocorrelation up to circa 60 pixels. In comparison, no significant positive spatial autocorrelation was found for the GWR model residuals at the distance more than 5 - 8 pixels. This suggests that the calibration of a local model reduces the problem of spatially autocorrelated error terms.

## 4 CONCLUSIONS

In this paper we revealed spatial variation in the spatial relationship between normalized difference vegetation index (NDVI) and rainfall in a sub-equatorial region of Central Sulawesi, Indonesia. The study investigated this variation both at and between spatial scales. The analysis based on the use of two different regression techniques: one is the global ordinary least squares regression, OLS, and the other is the geographically weighted regression, a relatively new local regression technique which allows the regression parameters and the strength of the relationship to vary over space. The analysis proved the presence of non-stationarity in the NDVI-precipitation relationship both between the main land-cover types and between locations. It means that the modelling of this relationship with the global or stratified OLS regression attains results with high amount of uncertainty. The variance in the relationship across the space of the study region is explained by the variance in the underlying environmental factors such as vegetation composition, soil type, hydrology, land use etc. caused by the diversity of terrain. That agrees with the results of recent studies on vegetation-climate relationships from other regions (Yang et al, 1998; Ji & Petters, 2004).

**Figure 6. Moran's I autocorrelation of regression residuals from the global OLS model (dashed line) and from the GWR model with a bandwidth of 1750 m (solid line). Obviously, that the residuals from the GWR model exhibit no significant autocorrelation at the distance more that 3-10 pixels, while the residuals from the OLS model are autocorrelated to a distance up to 60 pixels.**



Spatial non-stationarity of the relationship between NDVI and precipitation contributes essentially to scale-dependency in the results of the analysis (Foody, 2005). The GWR model enabled to use kernel bandwidth with different size (750 - 25000 m) working like some sort of a spatial microscope and scaling the modelling relationship from sub-regional to local scale and helping to determine the most appropriate scale. The results have shown that the regression parameters, the predictive power as well as the rank of the explanatory variable in the model of vegetation patterns is considered to represent a function of spatial scale. The results suggest that the explanatory power of the analysis increased very significantly with a diminishing of the scale. The NDVI-precipitation modelling provides the most accurate prediction by the use of the GWR model with a bandwidth of 1750 m. This model explains about 90 % of all variance in NDVI over the study area. Further decreasing of the analysis scale results in enlarge of AIC of the model and Moran’s *I* autocorrelation of its residuals. We will not discuss here the question “why” the model works best at this scale. Certainly, there should be hints from the climate research that could explain the spatial scale of the relationship. These hints will be thoroughly addressed in a paper that is being prepared by the authors and will be submitted shortly.

The results suggest that the calibration of local rather than global models reduces the problem of spatially autocorrelated errors. The residuals from the global OLS model clearly exhibited positive spatial autocorrelation up to approximately 60 pixels. In comparison to that, the residuals from the GWR model showed positive autocorrelation at the distance at least 10 times shorter, suggesting the ability of GWR approach to deal with spatial non-stationary problems. The GWR provides a more directly interpretable solution to the problem of spatially autocorrelated errors in spatial modeling compared with the global forms of spatial regression modelling. In GWR, the spatial non-stationarity of the parameters is modelled directly, rather than allowing the non-stationarity to be reflected through the error terms in the global model. This agrees with the results that have been discussed by Fotheringham et al. (2002) and Wang et al. (2005).

Our study proved the superiority of the local approach provided by GWR over the global OLS approach in analysing the relationship between patterns of NDVI and precipitation. This superiority is mainly due to the consideration of the spatial variation of the relationship over the study region. Global regression techniques like OLS may ignore local information and, therefore, indicate incorrectly that a large part of the variance in NDVI was unexplained. The non-stationary modelling based on the GWR approach has the potential for a more reliable prediction because the model is more aligned to local circumstances, although definitely a greater number of data is required to allow a reliable local fitting.

**ACKNOWLEDGMENTS**

This research work has been carried out within the interdisciplinary research project SFB-552 "STORMA" (Stability of Rainforest Margins in Sulawesi, Indonesia). The project is funded by the Deutsche Forschungsgemeinschaft (DFG).

**REFERENCES**

Asrar, G. M., Fuchs, M.m Kanemasu, E. T. & Hatfield, J. L. (1984). Estimating absorbed photosynthetically active radiation and leaf area index from spectral reflectance in wheat. Agronomy Journal, 87: 300-306.

Brunsdon, C., Fotheringham, A. S., Charlton, M. E. 1996. Geographically weighted regression: a method for exploring spatial non-stationarity. Geographical Analysis, 28: 281-298.

Chen, J., Jönsson, P., Tamura, M., Gu, Z., Matsushita, B., Eklundh, L. (2004). A simple method for reconstructing a high quality NDVI time series data set based on the Savitzky-Golay filter. *Rem. Sens. Environment,* 91: 332-344.

Chiles, J. P. and Delfiner, P. (1999). *Geostatistics: Modeling spatial uncertainty*. John Wiley & Sons, New York.

Clifford, P., Richardson, S. and Hemon, D. (1989). Assessing the significance of the correlation between two spatial processes. *Biometrics,* 45: 123-134.

Erasmi, S., Bothe, M. and Petta, R. A. (2006). Enhanced filtering of MODIS time series data for the analysis of desertification processes in northern Brazil. *The international Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences,* Vol. 36, Part 7.

Erasmi, S., Knieper, C., Twele, A. and Kappas, M. (2007): Monitoring inter-annual land cover dynamics at the rainforest margin in Central Sulawesi, Indonesia. In: Bochenek, S. (Eds.): New Developments and Challenges in Remote Sensing. Milpress, Rotterdam, pp. 297-308

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

Foody, G. M. (2004). Spatial nonstationary and scale-dependancy in the relationship between species richness and environmental determinants for the sub-Saharan endemic avifauna. *Global Ecology and Biogeography*, 13: 315-320.

Fotheringham, A. S., Brunsdon, C. and Charlton, M. (2002). *Geographically weighted regression: the analysis of spatially varying relationships*. Chichester, Willey.

Fotheringham, A. S., Charlton, M. E. and Brundson, C. (1996). The geography of parameter space: and investigation into spatial non-stationarity*.* *International Journal of GIS*, 10: 605-627.

Griffith, D. A. (2003). *Spatial autocorrelation and Spatial Filtering*. Berlin, Springer-Verlag.

Holben, B. N. (1986). Characteristics of maximum-value composite images from temporal AVHRR data. *International Journal of Remote Sensing*, 7:1417-1434.

Jetz, W., Rahbek, C. and Lichstein, J. W. (2005). Local and global approaches to spatial data analysis in ecology. *Global Ecology and Biogeography*, 14: 97-98.

Ji, L. and A. J. Peters. (2004). A Spatial Regression Procedure for Evaluating the Relationship between AVHRR-NDVI and Climate in the Nothern Great Plains. *International Journal of Remote Sensing*, 25: 297-311.

Justice, C. O., Townshend ,J. R. G., Holben, B. N., Tucker, C.J., (1985). Analysis of the phenology of global vegetation using meteorological satellite data. *Int. J. of Remote Sensing*, 6: 1271–1318.

Leung, Y., Mei, C. L., Zhang, W. X. 2000. Statistical tests for spatial non-stationarity based on the geographically weighted regression. Environment and Planning, A 32: 9-32.

Li, B., Tao, S. and Dawson, R. W. (2002). Relation between AVHRR NDVI and ecoclimatic parameters in China. *International Journal of Remote Sensing*, 23: 989-999.

Li, J., Lewis, J., Rowland, J., Tappan, G., Tieszen, L. (2004). Evaluation of land performance in Senegal using multi-temporal NDVI and rainfall series. *Journal of Arid Environments*, 59: 463-480.

Norcliffe, G. B. 1981. *Statistik für Geographen: eine Einführung*. Springer, Berlin.

Openshaw, S. (1984). *The modifiable areal unit problem*. CAT-MOG, 38. Geo Abstracts, Norwich.

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

Paez, A., Uchida, T. and Miyanmoto, K. (2002b). A general framework for estimation and inference of geographically weighted regression models: 2. Spatial association and model specification tests. *Environment and Planning A*, 34: 883-904.

Propastin, P. and M. Kappas. (2008). Reducing uncertainty in modelling NDVI-precipitation relationship: a comparative study using global and local regression techniques. *GIScience and Remote Sensing*, 45: 1-25.

Richard Y. and Poccard I. (1998). A statistical study of NDVI sensitivity to seasonal and inter-annual rainfall variations in southern Africa. *International Journal of Remote Sensing*, 19: 2907-2920.

Sellers, P., Randall, D. A., Betts, A. H., Hall, F. G., Berry, J. A., Collatz, G. J., Denning, A. S., Mooney, H. A., Nobre, C. A., Sato, N., Field, C. B. & Henderson-Sellers, A. (1997). Modelling the exchanges of energy water and carbon between continents and atmosphere. Science, 275: 502-509.

Tucker C. J. and P. J. Sellers. (1986). Satellite remote sensing of primary vegetation. *International Journal of Remote Sensing*, 7: 1395-1416.

Wang, J., Price, K. P. and P. M. Rich. (2001). Spatial patterns of NDVI in response to precipitation and temperature in the central Great Plains. *International Journal of Remote Sensing*, 22: 3827-3844.

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

Whitten, T., Henderson, G. S. and Mustafa, M. 2002. *The ecology of Sulawesi*. The ecology of Indonesia Series, 4, Jakarta.

Yang, L., Wylie, B., Tieszen, L. L., Reed, B. C., (1998). „An analysis of relationships among climate forcing and time-integrated NDVI of grasslands over the U.S. Northern and Central Great Plains. *Remote Sensing of the Environment*, 65: 25–37.