Study on the Method of Grass Yield Model in the Source Region of Three Rivers with Multivariate Data

Haoyan You[1,2] **, Chengfeng LUO**[1] **, Zhengjun Liu**[1] **, Jiao Wang**[1]

[1] Institute of Photogrammetry and Remote Sensing, Chinese Academy of Surveying and Mapping, Beijing 100830, China
[2] School of Geometrics, Liaoning Technical University, Fuxin, 123000, China

E-mail: youhaoyan1988@126.com, chfluo@casm.ac.cn, zjliu@casm.ac.cn, majestic5@126.com

Abstract. This paper uses remote sensing and GIS technology to analyse the Source Region of Three Rivers (SRTR) to establish a grass yield estimation model during 2010 with remote sensing data, meteorological data, grassland type data and ground measured data. Analysis of the correlation between ground measured data, vegetation index based HJ-1A/B satellite data, meteorological data and grassland type data were used to establish the grass yield model. The grass yield model was studied by several statistical methods, such as multiple linear regression and Geographically Weighted Regression (GWR). The model's precision was validated. Finally, the best model to estimate the grass yield of Maduo County in SRTR was contrasted with the TM degraded grassland interpretation image of Maduo County from 2009. The result shows that: (1) Comparing with the multiple linear regression model, the GWR model gave a much better fitting result with the quality of fit increasing significantly from less than 0.3 to more than 0.8; (2) The most sensitive factors affecting the grass yield in SRTR were precipitation from May to August and drought index from May to August. From calculation of the five vegetation indices, MSAVI fitted the best; (3) The Maduo County grass yield estimated by the optimal model was consistent with the TM degraded grassland interpretation image, the spatial distribution of grass yield in Maduo County for 2010 showed a "high south and low north" pattern.

1. Introduction

The grass yield which embodies grassland productivity is the basis for livestock husbandry production management. Estimating the grass yield through remote sensing image data has become important for grassland productivity research and grassland management. Biophysical models and the remote sensing data can be used as the drive to estimate the grass yield [1-4]. There are many studies which analyse the correlation between NDVI and grassland biomass, establishing an empirical model to estimate the yield [5-7]. Advantages and disadvantages of these two models are very obvious. Biophysical iological-physical models are used for large scale macro estimation, yet the estimation is not relatively accurate. With regards to reaching an accurate estimation of the local area, modeling work is very complex and has a close relationship with the natural characteristics of the regional area. Empirical models based on remote sensing data and ground measured data to build statistical models. They are simple and practicable. Yet, common models generally ignore the spatial heterogeneity of the

Haoyan You ,Email:youhaoyan1988@126.com

grass yield estimation and result in low accuracy. There are also many scholars who established remote sensing integrated models for estimating the grass yield[8]. To establish the model using remote sensing data, plant physiology, vegetation types, meteorological factors, soil texture, social data and other sources of information. There can be consideration of many factors. Relatively, the physical meaning is clear, the precision is higher, due to its acquisition difficulty in data and more restricted conditions, in practical applications it is difficult.

The geographically weighted regression(GWR)model, created the conditions for the spatial regression analysis. Ma Zongwen (2011) used this model to analyse the spatial relationship between NDVI and the natural and human factors of Bohai area[9]. Shao Yixi (2010) used this model to simulate regional land usage pattern[10]. The grass yield is related to geographical location, adjacent relations between geographical positions created the grass yield spatial correlation. The research uses the geographical weighted regression model and brought the spatial characteristics of grass yield into regression model for analysis. Combined with the ground measured data in SRTR and remote sensing data, optimal correlation factor of grassland yielded to estimate model, compare precisions between different parameter estimated models, and used the best model to estimate the grass yield of Maduo County in SRTR.

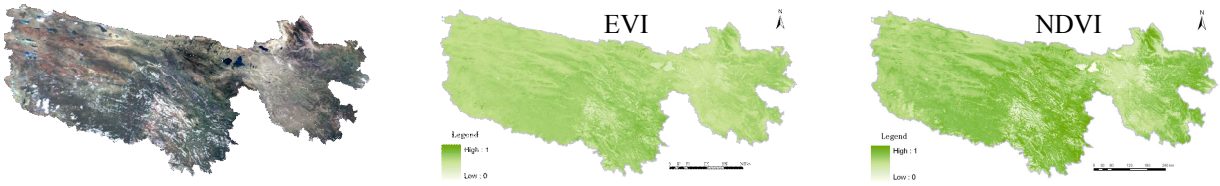
2. Study area and data

2.1. Study area

SRTR lies in western China, southern Qinghai Province, north latitude 31°39′~36°12 and longitude 89°45′~102°23. As the hinterland and main body of Qinghai-Tibet Plateau, SRTR is the biggest natural protection area in China. The total area is 346500 km^2 here, accounting for about 43% of the total area of Qinghai province. The mountain landscape is dominating here with the average elevation about 4400m and the landform is complex. The climate belongs to Tibetan Plateau climate system. Because of the high elevation, there is thin air in most regions and the growth period here is short. The main vegetation types are meadow complying with obvious horizontal distribution and vertical distribution rules. SRTR grassland mainly includes alpine meadow, alpine steppe and the smaller temperature steppe [11]. The alpine meadow generally grows in an elevation about 3500-4500m area and Kobresia is the dominant species. The alpine steppe usually located in the elevation about 4000-4500m area. Cold and xerophytic perennial dense grass and sedge are the dominant species. The temperature steppe is mainly distributed in valley beach and the Yellow River Valley in the northeast of the SRTR, where the dominant plants are long grass, northwest Stipa, Stipa breviflora, Achnatherum splendens and artemisia.

2.2. Data source and processing

The remote sensing data is the HJ-1A/B CCD data with resolution of 30 meters and the acquisition time is mid-August 2010. The image data preprocessing includes geometric precision correction, atmospheric correction based on 6S model and the SRTR image stitching and cutting. The calculation of the vegetation index about NDVI, EVI, SAVI, OSAVI, MSAVI[12] was achieved by the band operation.



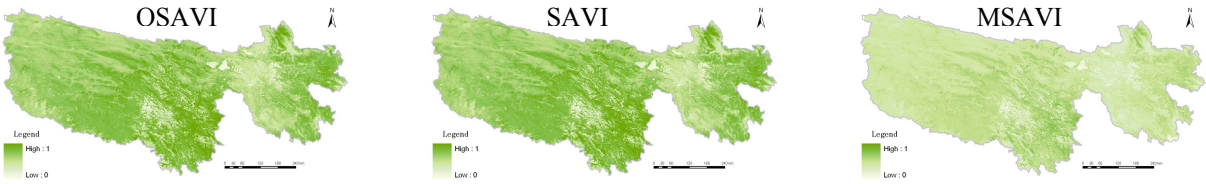


Figure 1. HJ data and the vegetation index data in SRTR 2010.08

The meteorological data is daily precipitation, temperature and the calculated drought index from the 20 stations in SRTR and its neighboring. Using ordinary Kriging interpolation on temperature, precipitation, drought index of the 20 sites, the meteorological data was derived in the SRTR as shown in the figure 3. The spatial distribution is consistent with grassland type data and the vegetation index data. The greatest impact on the grass yield level in growth season August is from the water and heat conditions in August of that year, the most growth summer (June to August), the growth period of spring (March to May) the green grass after May (May to August), spring to the growth season (March to August) in order to study the influence of meteorological conditions in different time on grass yield This study divided meteorological data into 5 time periods as previously mentioned andcalculated meteorological values for different period of time in SRTR.

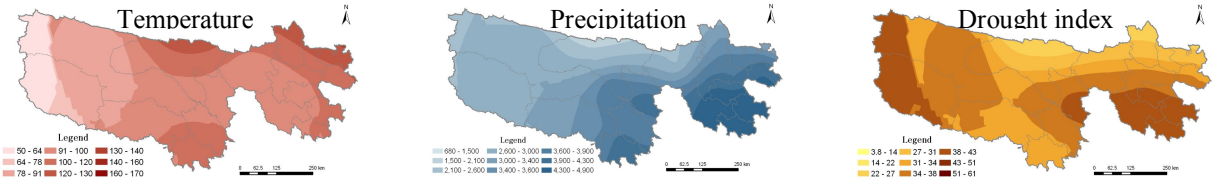


Figure 2. The meteorological data in SRTR 2010

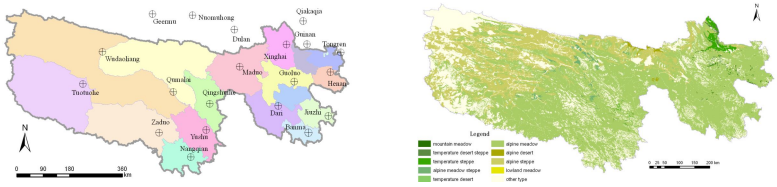


Figure 3. The meteorological stations distribution and grassland type data in SRTR 2010

The grassland type data contains 9 types in SRTR and quantifies the data on the basis of the productivity. The SRTR grass yield ground measured data acquainted in mid-August 2010.

3. Method

This study uses vegetation index, meteorological data, the grassland type data and ground measured data, through multiple linear regression and GWR to establish optimal grass yield estimation model of the SRTR.

GWR model is expressed as follows:

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i) x_{ik} + \varepsilon_i, i = 1,2,\cdots,n \tag{1}$$

Where($y_i; x_{i1}, x_{i2}, \cdots, x_{ip}$)is the observation value of dependent variable y and independent variable $x_1, x_2, \cdots, x_p$ in the position (u_i, v_i) ($i = 1,2,\cdots,n$). $\beta_j(u_i, v_j)(j = 0,1,\cdots,p)$ is the *ith* observation point unknown parameter in the position (u_i, v_i) ,and it is an arbitrary function of (u_i, v_i) , $\varepsilon_i(i = 1,2,\cdots,n)$ is error independent and identically distributed, generally are assumed to obey $N(0, \sigma^2)$. According to Tobler's first law of geography, the influence on the estimation of

parameters for point i from the observation value near point i on is greater than influence from that faraway from i, GWR model based on linear regression model assumes that the regression coefficient is an arbitrary function of observation point locationand it brings the spatial character into the model. The model calculates a local equation at each point An observation value weight is no longer remain constant in the regression process, and the weight is related to the proximity position to i:

$$\hat{\beta}(u_i, v_i) = (X^T W(u_i, v_i) X)^{-1} X^T W(u_i, v_i) Y \tag{2}$$

Where $\hat{\beta}$ is the estimation of β, the first row elements of the independent variable matrix are 1; X is the independent variable matrix about model factors; Y is the vector of the values of the dependent variable and it is the matrix about the grass yield ground measured data; W is a square matrix of weights relative to the position of (u_i, v_i) in the study area; in practics the calculation method of spatial weights matrices are the Gauss distance, exponential distance and tricube distance. This study chose Gauss distance to determine the weight:

$$W(u_i, v_i) = e^{-\frac{1}{2}(\frac{d_{(u_i, v_i)}}{b})^2} \tag{3}$$

where $W(u_i, v_i)$ is the geographical weight of the ith observation in the dataset relative to the location (u_i, v_i), $d_{(u_i, v_i)}$ is some measurement of the distance between the ith observation and the location (u_i, v_i), b is a quantity known as the bandwidth.

Bandwidth selection has a great influence on the operation results of GWR model. Bandwidth can be selected as a smoothing parameter, since the bigger bandwidth introduced bigger smoothing. As the bandwidth gets larger the weights approach unity and the local GWR model approaches the global OLS model. After several tests and analysis of test results, we use the bandwidth 230km and the corrected Akaike Information Criterion (Hurvich et al, 1998)[13] extensively in GWR for as the measurement of goodness of fit.

4. Model establishment and Precision verification

According to the ground measured data in 2010, 40 points were screened to establish the model. Extracted and measured correspond points are about the vegetation index data, meteorological data and grassland type data. The analysis of the correlation with the ground measured data is in order to understand the relationship of each factor with the ground measured data, establish the model through various regression methods, and verify the accuracy by the 15 reservation ground measured data.

4.1. Correlation analysis

The correlation analysis could reveal the relationship between two factors. In this study, the statistical analysis is used as the basis, with the coefficient of determination r^2 to determine the goodness of fit, analyze the correlation between grass yield with vegetation index, meteorological data, grassland type in SRTR, and choose the best factor to the model.

$$r^2 = \frac{[\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})]^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2} \tag{4}$$

Where n is samples number, X_i is independent variable (i=1,2,…,n), $\overline{X}$ is independent variable mean; Y_i are dependent variables (i=1,2,…,n), $\overline{Y}$ is dependent variable mean.

The determination coefficient r^2 determined the degree of relevance. When the r^2 is close to 1, the relevant equations are with high value reference; on the contrary, the closer to 0, the lower value reference.

The goodness of fit of GWR model increased significantly, from less than 0.2 to about 0.7, with the strongest correlation between ground measured data with MSAVI, precipitation from May to August and drought index from May to August, where r^2 of GWR MSAVI is 0.777, r^2 of GWR precipitation from May to August is 0.616, r^2 of GWR drought index from May to August is 0.617, and r^2 of GWR grassland type is 0.625.

4.2. Model establishment

This study chose the strongest correlation model factors MSAVI, the precipitation from May to August and drought index from May to August to establish the model. The ground measured data is chosen as the dependent variable and other factors are independent variables The software SAM is used to establish multiple linear regression model and GWR model. The multiple linear regression model is:

$$Y=467.06X_1-1.11\ X_2+\ 0.05\ X_3-76.88 \quad (5)$$

Table 1. The coefficient of linear regression

Variable	Coeff.	VIF	Std	Error
MSAVI	467.057	242.284	1.928	0.062
Drought index from May to August	-1.109	22.572	-0.049	0.961
Precipitation from May to August	0.052	0.106	0.491	0.626
Constant	-76.884	100.048	-0.768	0.447

The result of multiple linear regression shows that the effectiveness of the model factors in diffident region is invariant, which reflects the average level of the whole research. There are further analysis with GWR method.

Table 2. The range change of GWR variable

Variable	Minimum	Lwr Quartile	Median	Upr Quartile	Maximum
MSAVI	-1704.79298	-327.24191	178.36284	685.19349	2350.43645
Drought index from May to August	-65.64918	-25.80592	22.09998	49.92715	333.7041
Precipitation from May to August	-2.90652	-0.11547	-0.04385	0.25055	0.38839
Costant	-402.175	-141.51833	-113.34056	277.99679	3276.7122

Parameters in multivariate linear regression model are the same, as shown in table 2, the parameters of GWR model change with different geographical locations.

Table 3. The improvement of GWR

Source	r^2	AICc	Res.Sum.Sq.	D.F.	MS	F	P-value
OLS Residuals	0.198	497.514	473196.26	2.00			0.018
GWR Improvements	0.66	13.449	35341.44	8.65	40504.17606		
GWR Residuals	0.858	484.065	122854.82	29.35	4185.78489	9.6766	<0.001

Comparing the goodness of fit of different regression model, we found that in the GWR model, AICc (Akaike Information Criterion)value decreased significantly and the determination coefficient r^2 increased significantly. Through the GWR model, r^2 is 0.858 with the model factors MSAVI, precipitation from May to August and drought index from May to August. Based on this model, bringing grassland type data to GWR model, the coefficient of determination r^2 is 0.792, which is not the optimal model. Maybe the grassland type of ground measured data is not representative. Since the grassland type is consistent, this study established grass yield model without grassland type data.

4.3. Precision verification

Through previous analysis, the optimal model used the factors MSAVI, precipitation from May to August and drought index from May to August. The accuracy is verified by the 15 reservation ground measured data. The use of RMSE meets the model forecast value appraisal standard. The calculation formula is as following:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{(P_i - O_i)^2}{P_i}} \quad (5)$$

Theoretically, when RMSE=0 the model is the ideal state. Without error, in practical applications, when the predictive value is closer to the critical value, the fitting result is better and the accuracy of the estimation model is obtained by 1-RMSE.

After calculation, GWR model of factors MSAVI, precipitation from May to August and drought index from May to August accuracy is 71.71%.

5. The grass yield estimation result of Maduo in SRTR

Through the analysis, using the model of factors MSAVI, precipitation from May to August and drought index from May to August to establish GWR model, the fitting r^2 is highest and the verification accuracy meets the requirements. This model is used to estimate the grass yield of Maduo County in SRTR, and the contrast with the TM degraded grassland interpretation image of Maduo County in 2009. GWR model used 40 ground measured dataand each point has a regression equation. Calculating the distance from every pixel in the image to the 40 points could determine which equation to use. Finally the non grassland region was removed with grassland type data.

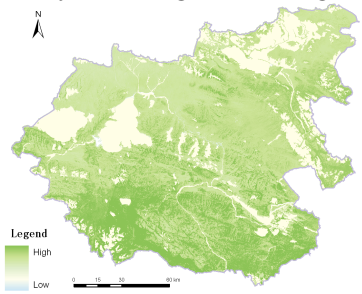


Figure 4. The grass yield of Maduo County in SRTR 2010.08

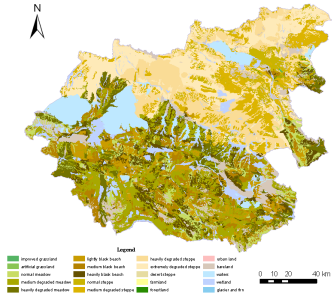


Figure 5. The TM degraded grassland interpretation image in SRTR 2009

The Maduo County grass yield estimations by the optimal model were consistent with the TM degraded grassland interpretation image. The spatial distribution of grass yield in Maduo County 2010 showed a “high south and low north” pattern.

6. Discussion and conclusion

This paper chose SRTR as a research area and used remote sensing, GIS technology to establish the grass yield estimation model in 2010 with remote sensing data, meteorological data, grassland type data and ground measured data. We tried to use GWR model to the grass yield monitoring, and compared with multiple linear regression model. The results showed that:

- Comparing with multiple linear regression model, GWR model gave a much better fitting result with the goodness of fit increased significantly from less than 0.3 to more than 0.8, and the accuracy raised from about 50% to more than 70%.
- Among the most sensitive factors affecting the grass yield in SRTR, which are mainly precipitation from May to August and drought index from May to August, and calculation of the 5 vegetation index, MSAVI was the best model fitting effect. Maybe this is because MSAVI eliminates the influence of soil to a certain extent.

- The Maduo County grass yield estimated by the optimal model was consistent with the TM degraded grassland interpretation image The spatial distribution of grass yield in Maduo County 2010 showed a “high south and low north” pattern.
- The bandwidth of GWR model must be appropriate.When the bandwidth is too large, due to the smoothing effect of large scale, model fitting becomes worse. When the bandwidth is too small, because the amount of data is too small, the computation can not be completed. In addition, the yield effect elements select not only comprehensive consideration, but also avoid the problem of multicollinearity among the elements.

Acknowledgements
This research is supported by National Natural Science Foundation of China (No.40901228), and the basic scientific research fund of Chinese Academy of Surveying and Mapping (No.7771301). The authors would like to thank Mr. Wang Y and Mr. Xu C (Qinghai Geomatics Center) for their useful suggestion.

References
[1] Bella C.DI., Faivre R., Ruget E, et al 2004 Remote sensing capabilities to estimate pasture production in France *International Journal of Remote Sensing*. **25** 5359-72
[2] Fan J , Zhong H, Harris W, et al 2008 Carbon storage in the grasslands of China based on field measurements of above-and below-ground biomass *Climatic Change*. **86** 375-96
[3] Cayrol P., Chehbouni A., Kergoat L., et al 2000 Grassland modeling and monitoring with SPOT-4 VEGETAT ION instrument during the 1997-1999 SALSA experiment *Agricultural and Forest Meteorology*. **105** 91-115
[4] Tumer DP, Ritts WD, Cohen WB, et al 2005 Site-level evaluation of satellite-based global terrestrial gross primary production and net primary production monitoring *Global Change Biology*. **11** 666-84
[5] Xu Bin, Yang Xiuchun, Tao Weiguo, et al 2007 Remote sensing monitoring upon the grass production in China *Acta Ecologica Sinica*. **2007** 405-13
[6] Yang Xiuchun, Xu Bin, Zhu Xiaohua,et al 2007 Models of grass production based on remote sensing monitoring in northern agro-grazing ecotone *Geographical Research*. **26** 213-22
[7] Lv Haiyan 2010 Study on the method of grass yield model: Chinese Academy of Agricultural Sciences
[8] Liu Jidong 2010 Study of rangeland degeneration base on climate herbage yield model and remote sensing herbage yield model: Agricultural University Of The Inner Mongol
[9] Ma Zongwen, Xu Xuegong, Lu Yaling 2011 Comparison of NDVI simulation models for Bohai Rim region and the factors affecting NDVI *Chinese Journal of Ecology*. **30** 1558-64
[10] Shao Yixi, Li Manchun, Chen Zhenjie, et al 2010 Simulation on regional spatial land use patterns using geographically weighted regression: A case study of Menghe Town, Changzhou *Scientia Geographica Sinica*. **30** 92-7
[11] Wang Kun, Hong Fuzeng, Zong Jinyao 2005 Resource resources and their sustainable utility in the "Three-River Headwaters" region *Acta Agrestia Sinica*. **13** 28-31
[12] Chen Pengfei, Wang Juanle, Liao Xiuying, et al 2010 Using data of HJ-1A/B satellite for hulunbeier grassland aboveground biomass estimation *Journal of natural Resources*. **7** 1122-31
[13] Hurvich CM, Simonoff JS and Tsai C9L 1998 Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion *Journal of Royal Statistical Societ y*. **B 60** 2719293