HERBACEOUS BIOMASS PREDICTION FROM ENVIRONMENTAL AND REMOTE SENSING INDICATORS

Nontembeko Dudeni-Tlhone[*1]. Abel Ramoelo[2]. Pravesh Debba[1]. Moses Azong Cho[2]. Renaud Mathieu[2].

[1]Decision Support and Systems Analysis, Spatial Planning Support, Built Environment Unit, Council for Scientific and Industrial Research (CSIR), P.O.Box 395, Pretoria, 0001, South Africa

[2]Earth Observation Research Group, Natural Resource and the Environment Unit, Council for Scientific and Industrial Research (CSIR), P.O.Box 395, Pretoria, 0001, South Africa

ndudenitlhone@csir.co.za

ABSTRACT

Feeding patterns and distribution of herbivores animals are known to be influenced by quality and quantity of forage such as grass. Modelling indicators of grass quality and biomass are critical in understanding such patterns and for decision makers such as park managers and farmers to efficiently plan and manage their rangelands. This study focused on predicting grass biomass using remote sensing and environmental variables. Since some of these variables were highly correlated, multivariate techniques such as partial least squares (PLS) and ridge regression were used to predict grass biomass in the Kruger National Park and the surrounding areas. The results indicated that both the environmental and remote sensing indicators had potential to predict grass biomass. Ridge regression showed better results since it explained about 41% of variation in the grass biomass, compared to the PLS model which explained approximately 33% variation.

1. INTRODUCTION

The health and quantity of rangeland resources such as grass are the primary drivers influencing the distribution and movement of herbivores (Drent and Prins, 1987; McNaughton, 1988; Ben-Shahar and Coe, 1992). Indicators of grass quantity and quality are particularly known to influence the feeding patterns of wildlife and livestock (Drent and Prins, 1987; McNaughton, 1988, 1990; Olff et al., 2002). Grass quantity is also referred to as biomass while quality often refers to foliar concentration of nutrients such as nitrogen, calcium and phosphorus. Techniques and platforms that can be used to model these key factors are essential

in developing in-depth understanding about the feeding patterns of herbivores (Ramoelo, 2012a). Such understanding is imperative for effective assessment, planning and management of ecological ecosystems.

Remote sensing platforms and various modelling or prediction techniques have previously been applied so as to improve knowledge about spatially-explicit resource distribution, quality and quantity. Some studies have been conducted to examine the role of climatic factors in grass production and quality (Thennissen, 1993; Skidmore et al 2011). It is often a challenge to model ecological ecosystems due to the dynamics and heterogeneity that exist in such ecosystems. It is therefore important that the modelling of such a complex ecosystem integrates with a wide range of factors that impact on grass biomass and health.

Studies have been undertaken to assess factors associated with grass quality (Knox et al., 2011, Ramoelo 2012a), and to our knowledge, little has been done to integrate remote sensing data and physical processes that affect grass quality and biomass. This study was undertaken to examine relevance of some statistical techniques in predicting grass quantity based on remote sensing indicators and environmental factors such as topography, climate, and land use.

2. METHODS

2.1. Study area

The study area is located in the north-eastern part of South Africa and covers Kruger National Park (KNP) and the soundings, including protected and privately-owned Sabi Sands Game Reserve (SGR), state-owned KNP, as well as communal land in the Bushbuckridge area. The private game reserve maintains various grazing activities by wild herbivores, namely, elephants, rhinos and impala. Meanwhile, communal rangelands support livestock grazing of mainly cattle, goats and sheep.

2.2. Data

The study area consisted of eight experimental sites across the land use gradient including the KNP, SGR and the communal areas. In order to capture grass biomass variability across these sites, transects were placed through various topographic surfaces such as valleys and crests. Forty nine (49) plots of about 30 m x 30 m were selected and subdivide into 3 to 4 subplots (0.5 m x 0.5 m), resulting in 189 individual sampling locations (sample size). The samples of the predominant grass species were then collected and dried at 80^0C for 24 hours. Each dried sample was weighed and measured in grams per square meter (g/m^2). The spatial locations of the samples were recorded using the Leica®'s differential geographic positioning system (DGPS).

Hyperspectral measurements of grass canopies that were taken using the Analytical Spectral Device (ASD) spectrometer; were used to derive the vegetation indices used in this study. The indices included the simple ratio (SR), the normalized difference vegetation index (NDVI) and the red edge position computed

through linear extrapolation (REPLI) (Cho et al., 2006). A few hyperspectral bands or absorption features associated with chlorophyll and nitrogen located in the different parts of the electromagnetic spectrum were also used. The nitrogen is known to influence biomass (Mutanga and Skidmore, 2004).

Topographic information including slope and aspect were computed from the digital elevation model (DEM) using ArcGIS software, while climatic indicators such as temperature and precipitation were sourced from the World Climatic database (www.worldclim.com). The geological data were obtained from the Council for Geoscience.

2.3. Prediction Models

Partial least squares (PLS) and ridge regression were considered in this study since, the aim was to predict grass biomass (dependent variable) from a number of predictor variables where some were collinear. Further, even though the relationships between these variables and biomass were not well understood, examining whether or not grass biomass could be predicted from the available data was critical in developing such understanding.

A general explanation of how these two techniques predict a response is given below.

A generalised form of a predictive model (least squares regression) can be defined as follows:

$$\mathbf{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \quad \varepsilon \sim N(0, \sigma^2),$$

where $\boldsymbol{Y}$ represents a matrix of dependent variables, $\boldsymbol{X}$ is a matrix of predictor variables and $\boldsymbol{\varepsilon}$ is a normally distributed error term. The least squares solution for the estimation of $\boldsymbol{\beta}$ is expressed as follows:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

If the number of variables in $\boldsymbol{X}$ exceeds the number of objects (sample size) or variables in $\boldsymbol{X}$ are highly correlated, then $\boldsymbol{X}^T\boldsymbol{X}$ becomes near singular or the inverse $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ does not exist or becomes unstable. This means that the estimation of $\boldsymbol{\beta}$ (which is known to be unbiased and has a minimum variance) becomes sensitive to a number of errors. These errors compromise the credibility of the prediction model.

In this situation, partial least squares solves this problem by extracting uncorrelated $\boldsymbol{T}$ and $\boldsymbol{P}$ scores and regressing $\boldsymbol{Y}$ on $\boldsymbol{T}$

1. $\boldsymbol{X} = \boldsymbol{TP}$
2. $\boldsymbol{y} = \boldsymbol{Tb}$ (Chong and Jun, 2005)

In equation 1 and 2, the $\boldsymbol{X}\ (n \times p)$, $\boldsymbol{T}\ (n \times h)$, $\boldsymbol{P}\ (p \times h)$, $\boldsymbol{y}\ (n \times 1)$, and $\boldsymbol{b}\ (h \times 1)$ are respectively used for predictors, $\boldsymbol{X}$- scores, $\boldsymbol{X}$- loadings, a response, and regression coefficients of $\boldsymbol{T}$. The *k*-th element of column vector $\boldsymbol{b}$ explains the relation between $\boldsymbol{y}$ and $\boldsymbol{t}_k$, the *k*-th column vector of $\boldsymbol{T}$.

Ridge regression, also known as regularization tries to address the problem of collinearity or excessive number of variables by adding a diagonal matrix $\lambda \boldsymbol{I}$ to $\boldsymbol{X^T X}$ where $\boldsymbol{I}$ is a $(p+1)$ by $(p+1)$ identity matrix. The +1 term in the $(p+1)$ by $(p+1)$ identity matrix is included to adjust for an intercept while no such term is included in the PLS, since PLS is performed on the centred matrices.

The ridge regression estimator can be denoted by the following expression.

$$\hat{\beta}_R = (\boldsymbol{X^T X})^{-1} + (\lambda \boldsymbol{I})^{-1}\ \boldsymbol{X^T Y}$$ (Hoerl and Kennard, 1970)

Where λ represents a penalization constant larger than zero. Lambda can be chosen by either using a cross-validation (CV) method or a generalized cross-validation (GCV) criterion, amongst other methods.

In short, ridge regression improves estimation of $\boldsymbol{\beta}$ by penalizing or standardizing the coefficients thereof and thereby minimizing its variance.

3. RESULTS AND DISCUSSION

Prior to showing and discussing the results from PLS and ridge regression methods which were used to predict grass biomass, it would be important to show existence of multi-collinearity using the standard regression method. The results from multiple regression showed that variables such as the spectral bands, NDVI, Altitude, Geology, and Landuse, had VIF values larger than a threshold of 10. Further, the correlation matrix for all the continuous predictors showed that all the bands were highly correlated with each other (correlation coefficients ranging from 0.84 to 0.87). Variables such as altitude, NDVI and precipitation were also highly correlated with coefficients in the range between 0.747 and 0.825. Stepwise regression analysis was also performed to address multi-collinearity and to assess the variables that had a significant contribution in the estimation of grass biomass. Altitude, geology and landuse were selected in the final stepwise regression, explaining about 34% of the variation in grass biomass.

This section continues with a summary of the prediction results from partial least squares and ridge regression methods. Partial least squares capabilities were also used to check for important variables in the prediction of grass biomass. This is particularly important when seeking to understand the role of the each of the explanatory variables in the prediction of grass biomass.

PLS Results

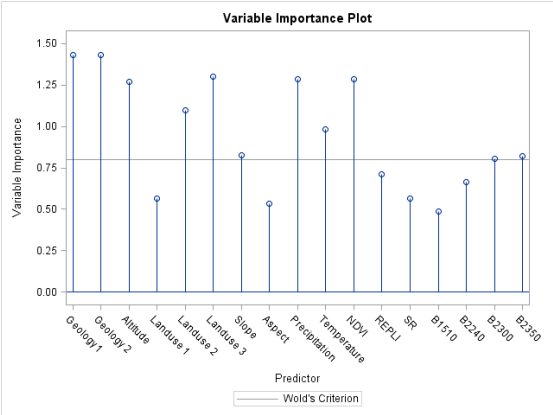

Figure 1: Variable very important projection (VIP) for prediction using PLS

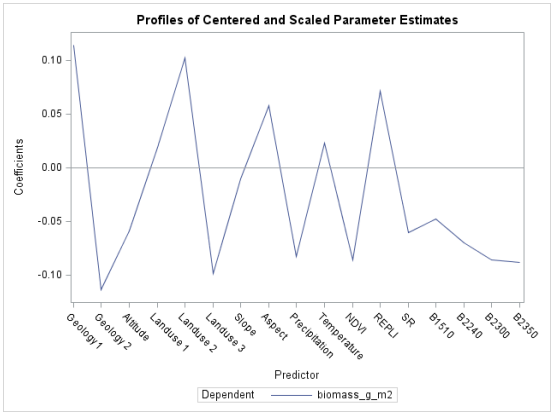

Figure 2: Absolute coefficients for predictor variables from PLS

Figure 1 shows all the variables used in predicting grass biomass and their relative importance in the PLS regression model. This is based on the statistic known as the very important projection (VIP) developed by (Wold, 1993), which measures the contribution of each predictor in the PLS model. A rule of thumb is used to evaluate the contribution of each variable in the model and to determine whether any variable could be a candidate for elimination or not. This rule implies that when a predictor has a Wold's value (VIP value) smaller than a threshold value of 0.8, that predictor could be considered for elimination.

A decision on whether to eliminate or keep the predictor in the model using VIP is, however, usually made in conjunction with (or validated by) the results from the centred and scaled parameter estimates statistic (shown in Figure 2). Essentially, if a predictor has a relatively smaller coefficient (in absolute value terms) and that coefficient is close to zero, and has a small VIP value (< 0.8), then that variable may be eliminated from the model.

According to the results shown in Figure 1, predictors including the geological classes (Geology1=granite soils and Geology2=grabbo soils), the land-use classes in the private reserves (Landuse2) and the communal areas (Landuse3), precipitation, temperature and NDVI appeared to be contributing more in the prediction of grass biomass since their VIP values were above the threshold of 0.8. Variables such as the land use class in the KNP (landuse1), aspect, and the remote sending indicators (REPLI, SR, spectral bands 1510 and 2240) were observed to contribute less in the prediction of grass quantity. Meanwhile variables like the slope and the spectral reflectance (B2300 and B2350) were approximately on the boarder mark (threshold region of 0.8) with regard to the Wold's value. This could indicate that the strength of these predictors is almost average in grass biomass estimation.

Figure 2 carries similar information to that contained in Figure 1 as it also indicates the degree of relationship (measured in terms of scaled coefficients) between the grass biomass and its predictors. The scaled coefficients for the predictors vary roughly between 0.02 and 0.10 (in absolute value terms) and this implies that the variables which are close to zero may have relatively smaller contribution in the prediction of grass quantity than those away from zero.

Figure 2 shows that temperature and aspect play a relatively small role in grass biomass estimation, with temperature having the smallest coefficient of all the variables. If we compare these results with those shown in Figure 1, we notice that for instance, temperature had a relatively larger VIP value and accordingly considered to have a contribution in grass biomass estimation while it was considered to be the least important in Figure 2. From this, it is difficult to make a decision on whether or not the temperature significantly contributes to the estimation of grass biomass. Further, we generally observed that these two figures did not maintain a clear and similar pattern in illustrating the contributory ability of each of the variables in the model. Hence, further investigation might be beneficial in understanding the contribution of each of the variable in this model.

Table 1: Summary of the correlation loadings from the PLS regression model

PLS factors	Predictor (X) R-square (%)	Response (Y) R-square (%)
1	30.30%	28.00%
2	24.60%	4.90%

Table 1 summarises the correlations between the predictors and grass biomass over the first two components. In the first component, the amount of variation explained by predictors in terms of the sum of squares is about 30% compared to that explained by grass biomass (28%).

Table 2: PLS analysis of grass biomass with cross-validation

	Factors		Responses		Cross-validation	
Number of PLS factors	Current	Total	Current	Total	Root mean PRESS	*P*
0					1.08	<.0001
1	30.29	30.29	28.01	28.01	0.93	0.04
2	24.63	54.92	4.89	32.90	0.91	0.19
3	7.64	62.56	3.54	36.43	0.90	0.26
4	2.82	65.38	2.11	38.54	0.88	0.82
5	3.42	68.80	1.01	39.55	0.90	0.15
6	8.38	77.17	0.60	40.14	0.89	0.55

Table 2 shows that the first six PLS factors account for approximately 40% of variation in the response variable (grass biomass) and over 77% of predictor variation. The absolute minimum PRESS is 0.88 obtained for 4 factors, but the model does not fit significantly better than the model with 2 factors. The 2 factor model is the simplest model with a satisfactory fit (p-value = 0.19) and should therefore be used in further analysis. This model (2 factor model) explained about 54% of the predictor variation and nearly 33% of the response variation.

Table 3: Parameter estimates from ridge regression

Variable	Parameter estimate	Pr > |t|
Intercept	-373.81	0.719

Table 4: Summary of the fit of the ridge regression model

	R-sq (%)	Adjusted R-sq (%)
Ridge regression	40. 98%	36.20%

Geology2	-54.81	0.000
Altitude	29.88	0.159
Landuse2	28.15	0.041
Landuse3	-25.27	0.991
Slope	21.00	0.024
Aspect	18.61	0.327
Precipitation	-31.91	0.305
Temperature	-6.98	0.233
NDVI	-39.77	0.366
REPLI	17.40	0.111
SR	-6.28	0.427
B1510	30.47	0.001
B2240	-16.73	0.080
B2300	-29.06	0.062
B2350	-24.36	0.960

In summary, Table 3 shows the results from the ridge regression where a ridge parameter of 14 estimated from the generalized cross-validation method was used. The results indicate that predictors such as a geological class representing the (grabbo soils), land use class in the private reserves (Sabi Sands) gradient, slope, and the spectral bands B1510, B2240 and B2300 were significantly important in the prediction of grass biomass.

Table 4 gives a summary of the results from ridge regression and indicates that the ridge regression model explained about 41% of the variation in grass biomass using the predictor variables listed in Table 2.

4. CONCLUSION

The objective of this study was to assess the potential of predicting grass biomass using environmental factors such as precipitation, aspect, temperature, land use, geology, altitude and slope as well as the remote sensing variables like NDVI, simple ratio, and some spectral bands. Partial least squares and ridge regression techniques were used to compute grass biomass predictions from these variables. The ridge regression explained nearly 41% of variation in the response compared to the partial least squares model which explained 32% of variation in the response. Even though some variables were deemed as being significant in the prediction of grass biomass, an understanding of how each variable contributed in the estimation of grass quantity was not developed. This challenge presents an opportunity for further research.

From these results, we can conclude that there is some potential in using the environmental factors and remote sensing information to predict grass biomass. In particular, topographic factors such as geology (granite and grabbo soils), land use classes, and slope could be very useful in assessing grass quantity.

REFERENCES

Ben-Shahar, R. & Coe, M. J. (1992). The relationships between soil factors, grass nutrients and the foraging behaviour of wildebeest and zebra. *Oecologia*, 90 (3), 422-428.

Cho, M., Skidmore, A. K., Corsi, F., van Wieren, S. and Sobhan, I., (2007). Estimation of green grass/herb biomass from airborne hyperspectral imagery using spectral indices and partial least squares regression. *International Journal of Applied Earth Observation and Geoinformation* 9(4): 414-424.

Chong, I.-G. & Jun, C.-H. (2005). Chemometrics and Intelligent Laboratory Systems 78, 103–112.

Drent, R. H. & Prins, H. H. T. (1987). The herbivore as prisoner of its food supply. In: Andel, J. V. & Bakker, J. (eds.) Disturbance in Grasslands: Species and Population Responses. Dordrecht: Dr. W. Junk Publishing Company.

Hoerl, A. E., Kennard, R. W. (1970). Ridge regression: Baised estimation for nonorthogonal problems. *Technometric*, 12: 55-67

Knox, N. M., Skidmore, A. K., Prins, H. H. T., Asner, G. P., Van Der Werff, H. M. A., De Boer, W. F., Van Der Waal, C., De Knegt, H. J., Kohi, E. M., Slotow, R. & Grant, R. C. (2011). Dry season mapping of savanna forage quality, using the hyperspectral Carnegie Airborne Observatory sensor. *Remote Sensing of Environment*, 115 (6), 1478-1488.

McNaughton, S. J. (1988). Mineral nutrition and spatial concentrations of African ungulates. *Nature*, 334, 343-345.

McNaughton, S. J. (1990). Mineral nutrition and seasonal movements of African migratory ungulates. *Nature*, 345, 613-615.

Mutanga, O. & Skidmore, A. K. (2004). Narrow band vegetation indices overcome the saturation problem in biomass estimation. *International Journal of Remote Sensing*, 25 (19), 3999 - 4014.

Mutanga, O., (2004). *Hyperspectral remote sensing of tropical grass quality and quantity*, Ph.D. thesis, ITC, Enschede.

Naes, T., Martens, H. (1985). Comparison of Prediction Methods for Multicollinear Data, *Communications in Statistics, Simulation and Computation*, 14, 545–576.

Olff, H., Ritchie, M. E. & Prins, H. H. T. (2002). Global environmental controls of diversity in large herbivores. *Nature*, 415 (6874), 901-904.

Ramoelo A. Skidmore A. K. Cho M. A. Schlerf M. Mathieu R. Heitkonig I. M. A. (2012b). Regional estimation of savanna grass nitrogen using the red-band of the spaceborne RapidEye sensor, *International Journal of Applied Earth Observation and Geoinformation*, 19, 151-162.

Ramoelo, A., (2012a). *Savanna grass quality: Remote sensing estimation from local to regional scale*, Ph.D. thesis, University of Twente.

Skidmore, A. K., Franklin, J., Dawson, T. P. & Pilejso, P. (2011). Geospatial tools address emerging issues in spatial ecology: a review and commentary on the Special Issue. *International Journal of Geographical Information Science*, 25 (3), 337-365.

Thennissen, J. D, (1993). Biomass production of different ecotypes of three grass species of the semi-arid grassland of Southern Africa, *Journal of Arid Environments*.

Tobias, R. D. Partial Least squares, SAS Institute Inc. Cary NC [Accessed online: 14 May 2012]

Wold, S. (1994). PLS for Multivariate Linear Modeling, *QSAR: Chemometric Methods in Molecular Design. Methods and Principles in Medicinal Chemistry.*

Wold, S., Johansson, E. & Cocchi, M. (1993). 3D QSAR in Drug Design; Theory, Methods, and Applications, ESCOM, Leiden, Holland, 523– 550.