SPATIAL OBJECT DEFINITION FOR VEGETATION PARAMETER ESTIMATION FROM HYMAP DATA

E. A. Addink *, S. M. de Jong , E. J. Pebesma

Department of Physical Geography, Faculty of Geosciences, Utrecht University, PO Box 80115, 3508 TC Utrecht, The Netherlands

KEY WORDS: Segmentation, scale, vegetation, HyMap, Ridge regression

ABSTRACT:

Common per-pixel estimations for vegetation parameters are hampered by spatial mismatch between the image and ground observations, and limited by neglecting spatial patterns. Geometric correction of images can reach accuracies in the range of 1 pixel, while locations of ground observations are measured with an accuracy of 5-10m by GPS. Our HyMap image has 5m pixels. Consequently, although coordinates may match, ground observations are not necessarily linked to the correct pixel, but can undesirably be represented by neighbouring pixels. Furthermore, vegetation patterns define the observation units used by ecologists, but they are not reflected by square pixels. Even though these patterns may reveal useful information, it is excluded from the analysis.

Object-oriented image analysis offers significant improvements. Objects are formed by groups of spectrally similar, neighbouring pixels; this reduces the risk of spatial mismatch. They are thus believed to provide a better approach to vegetation-parameter estimation than the conventional per-pixel approach.

Objects are defined by spectral similarity, but an important question is how much spectral variance is allowed. The aim of this paper is to investigate optimal heterogeneity for predicting biomass and LAI. We have data from 250 field plots in our test site, 60 km west of Montpellier in southern France. A HyMap image is available as well.

The image is segmented with different heterogeneities; larger heterogeneities resulting in larger segments. Field observations are linked to corresponding objects and with Ridge regression, relations between field observations and reflection values are identified. For each heterogeneity the prediction error is determined; the smallest error indicating optimal heterogeneity.

Conclusions confirm that increasing the object size shows an optimum in prediction accuracy for both biomass and LAI.

1. INTRODUCTION

1.1 General Instructions

Remote sensing data offer a powerful information source on vegetation parameters, which are needed in all sorts of models describing processes at the Earth's surface. Recently, hyperspectral data added even more power by providing spectral detail that allows detection of some chemical compounds of vegetation.

An even more recent development is object-oriented image analysis. Instead of analysing the spectral behaviour of individual pixels, neighbouring pixels are grouped either by segmentation or stratification. Segmentation is the process of grouping pixels based on spectral similarity, where maximum heterogeneity is the main parameter determining the result; shape constraints can be included as minor parameters. Larger heterogeneity values will result in larger segments. Stratification is the process of grouping pixels according to an external variable, the detail of which is fully determining the result.

Variables on hand in object-oriented image analysis are manifold that in per-pixel analysis. Beside the spectral variables (mean, sd) for each band, variables describing the shape and size are available, and a third group of variables describes relations with neighbouring objects.

Remote sensing classification studies profit enormously from this latter development. Object-oriented image analysis is much closer to human vision than the per-pixel analysis. Classification studies show both higher accuracy values and more detailed legends.

The application of object-oriented image analysis in vegetation parameter studies has been very limited so far, although it seems to offer improvements on two aspects. First, the geometric inaccuracies in both field and image data are of lesser importance, since a field plot is linked to an object rather than a pixel. The risk of linking it to a wrong object is much smaller than the risk of linking it to a wrong pixel, because of the larger spatial extent of objects. Secondly, field plots are often chosen such that they represent a vegetation patch. With per-pixel analysis, this information is ignored, while by grouping pixels, vegetation patches can show up (depending on the heterogeneity threshold).

In object-oriented-image-analysis literature very little attention is paid to optimal object definition. However, the definition is thought to affect the relations that are found between field observations on vegetation parameters and spectral information. Object definition comprises both the spectral bands included and the heterogeneity factor.

Furthermore, optimal object definition need not be identical for different vegetation parameters. For example, biomass and Leaf Area Index (LAI) are subjected to different dynamics. Biomass is determined by the accumulation of yearly growth. On the other hand, LAI is largely determined by the yearly situation; in the case of evergreens the situation of 2-4 years will determine it. Given this difference in temporal steering dynamics, the optimal object definition might just as well be different.

This paper focuses on the spatial aspects of object definition for vegetation parameter estimation, i.e. on the effect that the

* Corresponding author.

maximum heterogeneity has. Although the spectral bands definitely will have their effect on object definition as well, this will not be considered here. The aim of this paper is to answer two questions. 1). How does the spatial definition of objects affect the statistical relationships between field observations and spectral object properties? 2). Is this effect similar for different vegetation parameters? This paper will address those questions for a given data set; the validity for other areas will be explored in subsequent studies.
The optimal spatial definition is here defined as the level of segmentation that results in the lowest prediction error of the vegetation parameters.

2. DATA

2.1 Vegetation data

From August to October 2005 a field campaign was held in the La Peyne catchment in southern France, 60 km west of Montpellier. The vegetation in the study area frequently suffers from water and heat stress, as in all Mediterranean areas. It ranges from open areas with low herbal vegetation, *garrigue*, through dense bushes up to 5m, *maquis*, to the climax vegetation of the region, oak forests (Sluiter, 2005). Within the area some 250 plots were visited. Each plot measured 5m x 5m and was sampled for biomass and LAI.

Biomass: Biomass was estimated using empirical allometric formulas from Ogaya et al. (2003) for trees:

$$\ln AB = 4.900 + 2.277 \ln D50 \quad (1)$$

$$\ln AB = 3.830 + 2.563 \ln D50 \quad (2)$$

where AB = aboveground biomass (in Mg ha-1)
D50 = stem diameter at 50cm (in cm).
Equations (1) and (2) relate to the evergreen oak (*Quercus ilex*) and the strawberry tree (*Arbutus unedo*), respectively.
For shrubs we used a similar formula provided by Pereira et al. (1994):

$$AB = 0.642 \cdot H^{0.0075} \cdot D\max^{2.4901} \quad (3)$$

Where AB = aboveground biomass (in kg)
dmax = maximal diameter (in m)
h = height (in m).
Those equations were used to estimate the aboveground biomass for individual trees or shrubs, by summing all results per plot and dividing it by the plot area, values were transformed into the amount of biomass per hectare.
Biomass data were collected for 216 plots (table 1).

Leaf Area Index: In 243 plots photographs were taken with a hemispherical lens from below the canopy (oriented towards zenith) to estimate Leaf Area Index. Four photos were taken 1m from the corners on the diagonals, and one was taken in the centre of the plot. The photos were then analyzed with CAN-EYE (Baret and Weiss, 2004). This process consists of two steps: 1. photos are classified into one of two classes, vegetation or sky, 2. gap distribution is determined for different viewing angles. Jonckheere et al. (2004) and Weiss et al. (2004) give accurate descriptions of the underlying principles.
Leaf Area Index data were collected for 243 plots (table 1).

	Biomass	LAI
	Mg/ha	-
N	210	243
mean	167	3.2
sd	209	0.84
variance	43742	0.7
min	0.1	0.4
max	1347	5.4

Table 1. Statistical characteristics of field data.

2.2 Image data

A HyMap image recorded on 13 July 2003 covers the catchment of the Peyne river. The image has 124 bands and provides continuous spectral cover from 400 to 2500nm. The spatial resolution is 5m. The image was geometrically rectified using ground control points determined by GPS in the field and a 25m resolution DEM.

3. METHODS

3.1 Data processing

Masking: Since we were interested in vegetation parameters, only the vegetated pixels were included in the analysis to assure that spectral variance in the image was a result of variance in vegetation characteristics only. To remove the non-vegetated pixels a mask was produced in two steps. First all pixels with an NDVI value of 0.25 and less were selected. Next a buffer operation was applied, and the selected pixels were all buffered by two more pixels. Without the buffer, the pixels next to (masked-out) roads would show extreme values in the succeeding MNF transformation, indicating that they were affected by the neighbouring non-vegetated pixels.

Minimum Noise Fraction: The masked HyMap image, now only containing vegetated pixels, was subjected to a Minimum Noise Fraction (MNF) transformation (Green et al., 1988). With MNF noise reduction is applied by optimizing autocorrelation for an indicated part of the image. The remaining signal is then subjected to a Principal Component transformation. We applied MNF for two reasons. First, the number of bands (124) is very large to include in the segmentation procedure. Secondly, hyperspectral data show a high level of collinearity, i.e. correlation between the variables. Application of MNF allows reduction of the number of variables while maintaining most of the variance. Furthermore, it results in non-correlated bands, so it solves the collinearity problem.
With MNF the number of input and output bands is equal, the output bands showing decreasing variance. The analysis was continued with the first 20 MNF bands, which explained 84% of the total variance in the masked image.

Segmentation: Segmentation of the image was performed with eCognition 3.0, an object-based image analysis package (Definiens, 2003). Segments were exclusively defined by MNF values without any limitations from shape parameters. Within eCognition the maximal heterogeneity of the objects is set by the *scale parameter*. The MNF image was segmented ten times with increments of the scale parameter of 5. The exact

definition of the scale parameter is not published by Definiens. The number of segments decreased rapidly with increasing scale parameter values (Table 2).

Scale	N
5	8763
10	3752
15	2800
20	2458
25	2286
30	2203
35	2147
40	2103
50	2067

Table 2. Number of segments (N) resulting from segmentation with different scale parameters (Scale).

For each segment the mean value for each of the 124 HyMap bands was calculated.

Data set preparation: For both vegetation parameters 11 data sets were prepared, ten for the different segmentation levels, and one relating the field plots to individual pixels. So in total 22 data sets were prepared. Each data set, contained the parameter values and the 124 spectral band values. The MNF bands were only used to segment the images, while the relationship between the HyMap image and the vegetation data will be based on the original bands.

3.2 Statistical analysis

Ridge regression: The relation between spectral behaviour of vegetation and biomass and LAI is determined using Ridge regression (Hastie et al., 2001). This is a linear multiple regression method, which searches for the minimum of squared errors, while at the same time limiting the range of the squared sums of the regression coefficients. In situations with many correlated variables, like in hyperspectral images, regression coefficients become poorly determined and exhibit high variance. By imposing a size constraint on the coefficients this phenomenon is prevented.
The size constraint of the regression coefficients is set by λ. There is an inverse, non-linear relation between λ and the degrees of freedom (df) of the regression coefficients. With λ equal to 0, Ridge regression is equal to regular multiple regression with maximum df. By increasing λ, df will decrease (p63, eq. 3.50, Hastie et al., 2001).

Cross validation: The performance of the Ridge regression functions was determined using generalised cross validation (GCV). GCV values are calculated for each data set for the same range of λ values. GCV is equal to the total residual variance, so lower GCV values indicate better performance.

4. RESULTS

The results of the cross validation are given in Figures 1 and 2 for Leaf Area Index and Biomass, respectively. For both parameters ten graphs are provided, showing, from top-left to lower-right, the results for individual pixels to scale parameter 50. For lay-out reasons, the graphs for scale parameter 45 are not shown. However, in both cases these curves do confirm the trend shown by scale parameters 40 and 50.
The vertical axis shows the GCV values, while the horizontal axis shows the degrees of freedom, df. The lowest points of the graphs indicate the best performance for a given scale parameter. For Leaf Area Index, scale parameter 15 shows the lowest GCV minimum of the ten graphs. The GCV value of 0.38 corresponds to 54% unexplained variance, which means an R^2 of 0.46. For biomass, scale parameter 10 shows the lowest GCV value of the ten graphs. The value of 23000 results in an R^2 of 0.47.
For biomass optimal performance increases from individual pixels to scale parameter 10, after which it decreases again with larger scale parameters. For Leaf Area Index the initial trend is not so straightforward, with scale parameter 5 showing better results than scale parameter 10, although the optimum at scale parameters 15 or 20 is clearly better. From 25 on, the performance shows a clear decreasing performance.
Total variance in the data set determines the relation between the size constraint of the regression coefficients λ, and df. This shows in the smaller range covered by df with increasing scale parameters.

5. DISCUSSION

The different levels of segmentation result in different accuracy values for estimation of Leaf Area Index and biomass. Segmentation compared to the one-pixel situation shows that segmentation indeed does provide better estimates.
By segmenting the images, information is lost. Up to a certain level this is expected to be noise, stemming either from spectral noise or spatial mismatch. At a certain aggregation level the lost information might turn out to be relevant. This would show in worse results, in our case lower GCV values. Both phenomena can be observed in the GCV curves for the different scale parameters. Predictions improved until scale parameter 10 (biomass) or 15 (LAI), and decreased with subsequent scale parameter values.
This study does not aim at determining the exact scale parameters that yield optimal predictions, but merely at showing that different values yield different results. The optimal scale parameter can be derived by varying it with a smaller step size.

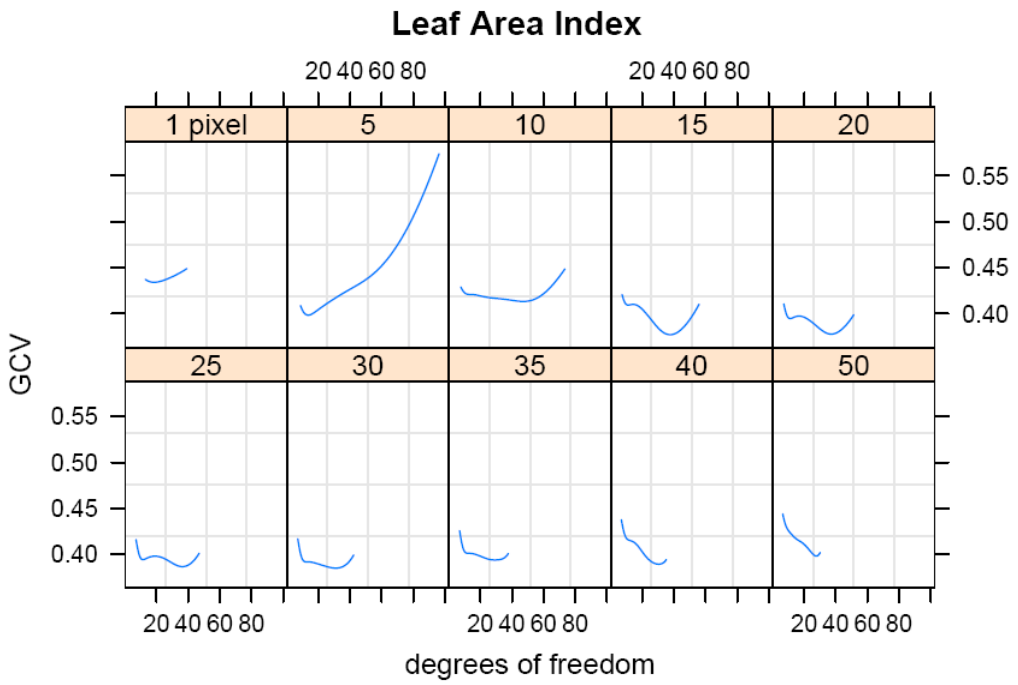

Figure 1. The Generalized Cross Validation (GCV) of Leaf Area Index plotted against the degrees of freedom for 10 different object definitions. GCV is equal to the unexplained variance, lower values indicate better estimates. Each plot from upper left to lower right corresponds to the scale parameters of table 2.

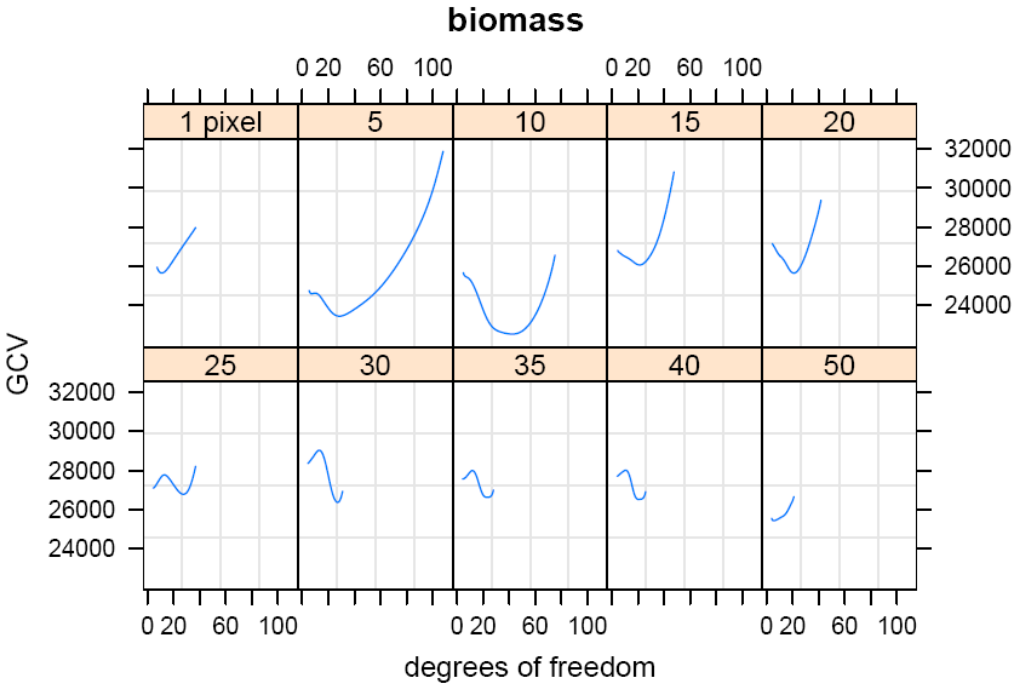

Figure 2. The Generalized Cross Validation (GCV) of biomass plotted against the degrees of freedom for 10 different object definitions. GCV is equal to the unexplained variance, lower values indicate better estimates. Each plot from upper left to lower right corresponds to the scale parameters of table 2.

The band setting for the segmentation was remained constant in this study by using the first 20 MNF bands for all segmentations. Finding optimal band combinations was beyond the scope of this paper.

6. CONCLUSIONS

In this paper we studied the effect of increasing heterogeneity in object definition on the accuracy of predicting biomass and Leaf Area Index values. We used Ridge regression to establish equations and by a leave-one-out cross validation the accuracies of the estimations were determined. We aimed at answering two questions; 1). How does the spatial definition of objects affect the statistical relationships between field observations and spectral object properties? 2). Is this effect similar for different vegetation parameters?

It can be concluded that different heterogeneities indeed result in different estimation accuracy values. Starting with individual pixels and increasing the object size, the predictions improve until an optimum is reached, after which increasing object size results in worse predictions. The question what determines the optimal setting is the next issue to study.

Furthermore, our results show that Leaf Area Index and aboveground biomass show different optima for their predictions. The explanation for this might be well related to the question of the underlying principles of optimal heterogeneity values.

REFERENCES

Baret, F. and Weiss, M., 2004. CAN_EYE : Processing digital photographs for canopy structure characterization. Tutorial. INRA Avignon. http://www.avignon.inra.fr/can_eye/.

Definiens, 2003. eCognition, object-based image analysis. München.

Green, A. A., Berman, M., Switzer, P. and Craig, M. D., 1988. Transformation for ordering multispectral data in terms of image quality with implications for noise removal. *IEEE Transactions on Geoscience and Remote Sensing* 26 (1), pp. 65-74.

Hastie, T., Tibshirani, R. and Friedman, J., 2001. *The elements of statistical learning. Data mining, inference, and prediction*. Springer, New York, pp. 59-64.

Jonckheere, I., Fleck, S., Nackaerts, K., Muysa, B., Coppin, P., Weiss, M. and Baret, F., 2004. Review of methods for in situ leaf area index determination Part I. Theories, sensors and hemispherical photography. *Agricultural and Forest Meteorology* 121, pp. 19–35.

Ogaya, R., Peñuelas, J., Martínez-Vilalta, J., and Mangirón, M., 2003. Effect of drought on diameter increment of *Quercus ilex, Phillyrea latifolia*, and *Arbutus unedo* in a holm oak forest of NE Spain. *Forest Ecology and Management* 180, pp. 175-184.

Pereira, J. M. C., Oliveira, T. M., and Paul, J. P. C., 1994. Fuel mapping in a Mediterranean shrubland using Landsat TM imagery. In: P. J. Kennedy and M. Karteris (eds.), *International workshop on satellite technology and GIS for Mediterranean forest mapping and fire management*, Office for official publication of the European Communities, Luxembourg, pp. 97-106.

Sluiter, R., 2005. *Mediterranean land cover change - Modelling and monitoring natural vegetation using GIS and remote sensing*. NGS Studies 333, KNAG, Utrecht, 145 pp.

Weiss, M., Baret, F., Smith, G. J., Jonckheere, I., and Coppin, P., 2004. Review of methods for in situ leaf area index determination Part II. Estimation of LAI, errors and sampling. *Agricultural and Forest Meteorology* 121, pp. 37-53.

ACKNOWLEDGEMENTS

Wiebe Nijland, Rogier de Jong and Paul Hiemstra are greatly acknowledged for their large contribution to the field campaign.