# AN INTEGRATED METHOD FOR FOREST CANOPY COVER MAPPING USING LANDSAT ETM+ IMAGERY

**Zhongwu Wang**, Remote Sensing Analyst
**Andrew Brenner**, General Manager
Sanborn Map Company
455 E. Eisenhower Parkway, Suite 240 Ann Arbor, MI 48108
zwang@sanborn.com
abrenner@sanborn.com

## ABSTRACT

Forest canopy cover is the proportion of the forest floor covered by the vertical projection of tree crowns. It is of great interest for a variety of land and fire management applications, many of which require not only information on land cover classification, but also accurate canopy cover information. In this paper a new method is developed for deriving canopy cover information using multispectral Landsat ETM+ imagery. The method integrates image segmentation and support vector machine (SVM) techniques to produce canopy cover data. Image segmentation can not only remove the effects of noisy pixels, but also generate a suite of spectral and spatial variables for SVM model. Historical canopy cover data interpreted from imagery of Google Maps, together with in situ field data were used for creating training samples for training and tuning the SVM regression model. The SVM regression model adopted in this study is e1071 package which is implemented in R software. R is a free software environment for statistical computing and graphics, and has great capability for spatial analysis. The method was tested for canopy cover for the State of Florida. This model demonstrated an efficient and accurate method of estimating canopy cover as an alternative to linear regression and regression tree methods.

Key words: Support vector machine, forest canopy cover, e1071 package, software R, remote sensing

## INTRODUCTION

Forest canopy cover is of great interest to a variety of scientific and land management applications, many of whom require not only information on forest categories, but also tree canopy density (Huang, 2001). Forest canopy cover, also known as canopy density, canopy coverage or crown cover, is defined as the proportion of the forest floor covered by the vertical projection of the tree crowns (Jennings, 1999). Estimation of forest canopy cover has become an important part of forest inventories, and remote sensing data have provided a useful and highly efficient way to estimate canopy cover. A plethora of different techniques have been devised to map forest canopy cover using remotely sensed data in the past, and linear regression model and classification and regression tree (CART) model are two of the most widely used methods (Larsson, 2000; Carreiras, 2006; Huang, 2001).

### Support Vector Machines

This study presents an alternative way to estimate forest canopy cover by using Support Vector Machine (SVM) approaches. Some people believe that SVM is the best "off the shelf" machine learning method. Support Vector learning is based on simple ideas which originated in statistical learning theory (Vapnik, 1998). The simplicity comes from the fact that SVM applies a simple linear method to the data but in a high-dimensional feature space non-linearly related to the input space (Karatzoglou, 2006). SVM technique is based on the idea of separating data with the large "gap", and it turns out that functional and geometric margin classifiers can be used to evaluate the "gap" in the data (shown in Figure 1). To solve SVM program is to find the solution to maximize the margin classifiers. Since it is difficult to solve the margin classifiers program directly, the form of margin classifiers has been converted to the form of Lagrange duality in order to find the optimization solutions. Then the problem is a standard convex optimization problem and can be solved in a quadratic programming package. Kernel trick techniques can also be used to map data into higher dimensional feature space without increasing the complexity of SVM problem.

SVM technique can also be applied to regression problems where the target variable is a continuous variable. This method is called support vector regression (SVR) and it is an extension to SVM classification model. The model produced by support vector classification (as described above) only depends on a subset of the training data, because the cost function for building the model does not care about training points that lie beyond the margin. Analogously, the model produced by SVR only depends on a subset of the training data, because the cost function for building the model ignores any training data that are close to the model prediction (Drucker, 1996).

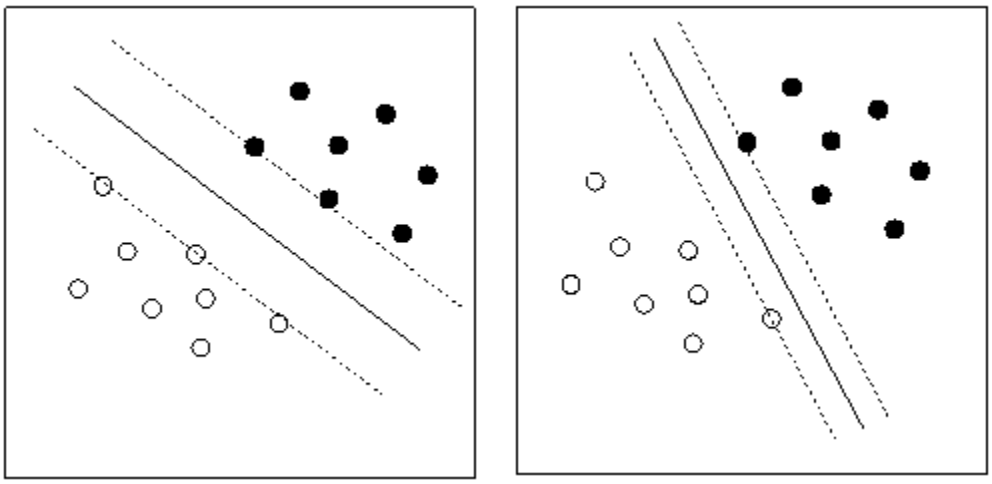

**Figure 1.** Comparison of linear SVM classifier with different marginal size. The left image has larger margin compared with the right one. The left is a better margin classifier solution.
(*From: http://mi.eng.cam.ac.uk/~kkc21/thesis_main/node12.html*)

SVM has been applied within the remote sensing community to multispectral and hyperspectral imagery analysis. For instance, Nemmour (2006) has used multiple support vector machines for land cover change detection and found that SVM is an efficient method for land cover change detection when compared with neural networks. Zhang (2008) has used an improved SVM method for remotely sensed image classification, and found that SVM method is very competitive in term of accuracy of classification of remotely sensed data and the time needed is less.

## METHOD

Gonzales (2004) has classified image analysis techniques into three categories: low-, mid-, and high-level image analysis. Low-level techniques involve primitive operations such as all kinds of image transformations like geometric correction. It is characterized by the fact that both inputs and outputs are images. Mid-level techniques involve tasks such as image segmentation (partitioning an image into non-overlapping objects), description of those objects to reduce them to a form (database) for computer processing, and classification of individual objects. This level is characterized by the fact that its inputs are general images, but its outputs are attributes extracted from those images. Finally higher-level image analysis involves “making sense of” the image objects. It is characterized by the fact that its outputs are pattern, knowledge or rules (classification, regression, association rules) hidden in the database. This study involved image processes of all those three levels. The first level was Tasseled Cap transformation and creation of NDVI image. The second level was segmentation of the input images and organization of those objects into an image object database. The last level was the use of training samples to generate regression rules through SVR model and then apply those rules to all image objects. Figure 2 illustrates the flow chart of image processing in this study.

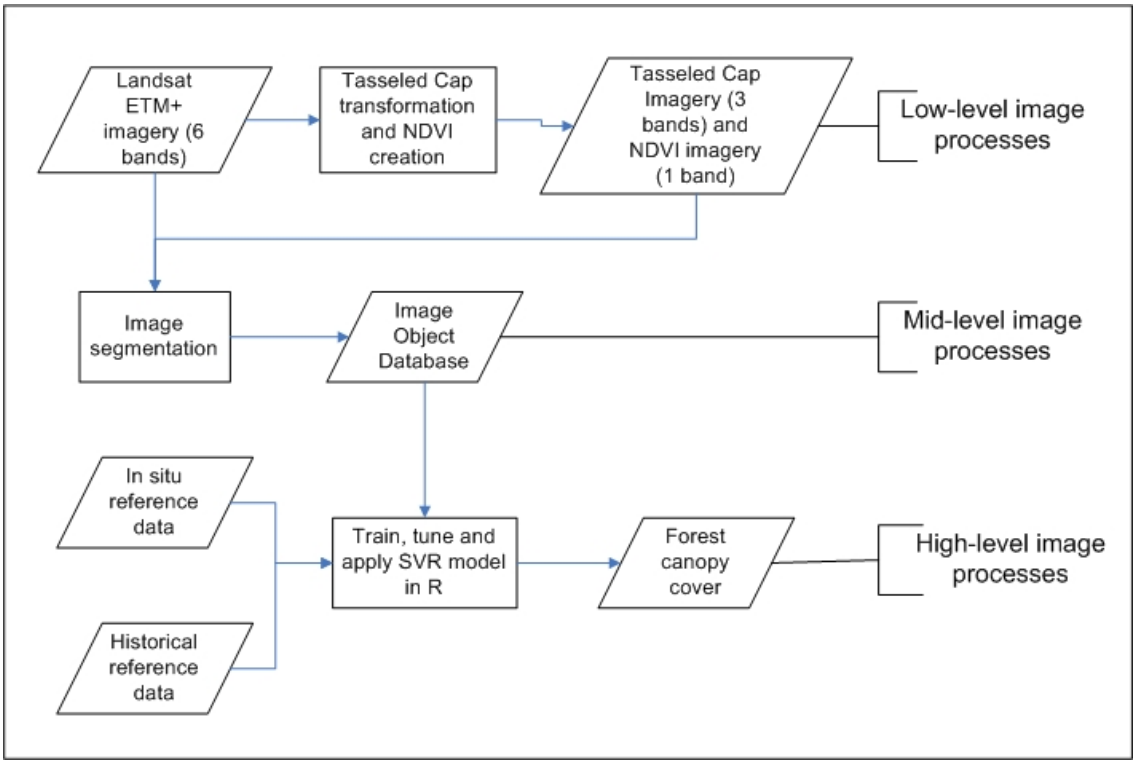


**Figure 2.** A flowchart of the strategy for deriving 30 m forest canopy data.

## Study Area and Remote Sensing Data

In this test, the SVR method was applied to Landsat ETM+ multispectral data acquired on March 25, 2007. The study area was in the panhandle area of the State of Florida. Part of the study area is Apalachicola National Forest which has high percentage of canopy cover. The original image had six reflectance bands; the spatial resolution was 30 m, and the image had 7027 x 6320 pixels. Figure 3 shows the study area.

The image was used to create a Tasseled Cap image with three spectral bands through Tasseled Cap transformation and an NDVI image with one spectral band.

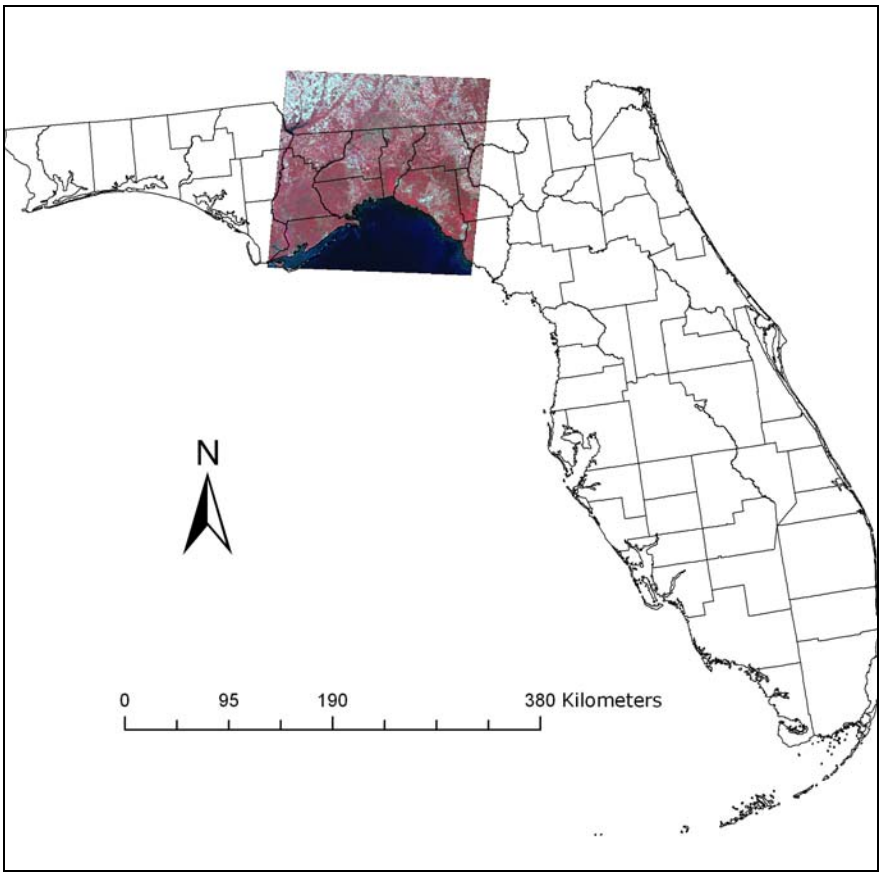


**Figure 3.** Study area of this experiment is located in Northwest of Florida. (Landsat ETM+ Image is display as RGB = 4, 3, 2).

## Image Segmentation and Image Object Database

The image with 10 spectral bands (6 reflectance bands, 3 Tasseled Cap bands, 1 NDVI band) was segmented using the Definiens algorithm with the parameter settings shown in Table 1. The most important parameter is the scale parameter, which is used to control the average segment size. Several different scale values were tested and the value of 20 was selected based on visual inspection of the resultant segments. The procedure used the parameters shown in Table 1. The Definiens segmentation resulted in 104,416 segments.

The segmentation allowed for the creation of an image object database. The image objects were the distinct features delineated in the segmentation process. Working with image objects can minimize noise related to the collection angles, illumination differences by averaging pixel values over an area, and also allows for collecting summary of derived objects (such as area, rectangular fit, and length). The resultant image object database had collected following information about image objects:

- 10 mean values of objects (each associated with one input spectral band)
- brightness value of objects which is an overall intensity measurement of all spectral layers (mean value of all reflectance layers).
- spatial information about objects including area, length and rectangular fit.

**Table 1.** Image segmentation settings for multi-resolution image segmentation in the Definiens software

| Scale Parameter | Color Criterion | Shape Criterion | |
|---|---|---|---|
| | | Compactness | Smoothness |
| 20 | 90% | 2% | 8% |

## Reference Data Collection

In situ data were collected by Sanborn (Ann Arbor, Michigan) and Landmark Systems (Tallahassee, Florida) in August 2008. In situ data documented detail information about geographic location, canopy cover and species for each plot. There were 683 field plots collected and most of the plots are located about 200 meters buffer around the road networks. There are two types of potential errors existing in statistical learning theory. One of them is approximation error where the sample space of training data is not representative of target space. To minimize the approximation error, more reference data needed to be collected away from the road networks. To achieve this objective, imagery was used to collect more reference data by photo interpretation. The approach was to create a land cover change mask and apply the mask to unchanged areas of a canopy cover dataset produced previously. The unchanged canopy cover portion was then stratified into 20 classes (each class representing a 5% canopy cover interval), and converted into polygons. Then 10-20 polygons were randomly selected from each of the 20 classes and used as potential training polygons with their canopy covers verified with imagery from Google Map, this ensured that the training sites were homogenous and had the correct canopy cover. An ArcObjects tool was created that geographically linked potential polygon to Google Map to facilitate the process. Figure 4 shows the tool.

One fifth of reference data were held independently as test data in the model training process to tune SVR model and another one fifth dataset were held independently for evaluating the accuracy of predicted canopy cover. The selection of test dataset was stratified with canopy cover value uniformly distributed over the cover classes. The rest of the dataset was used to train SVR model.

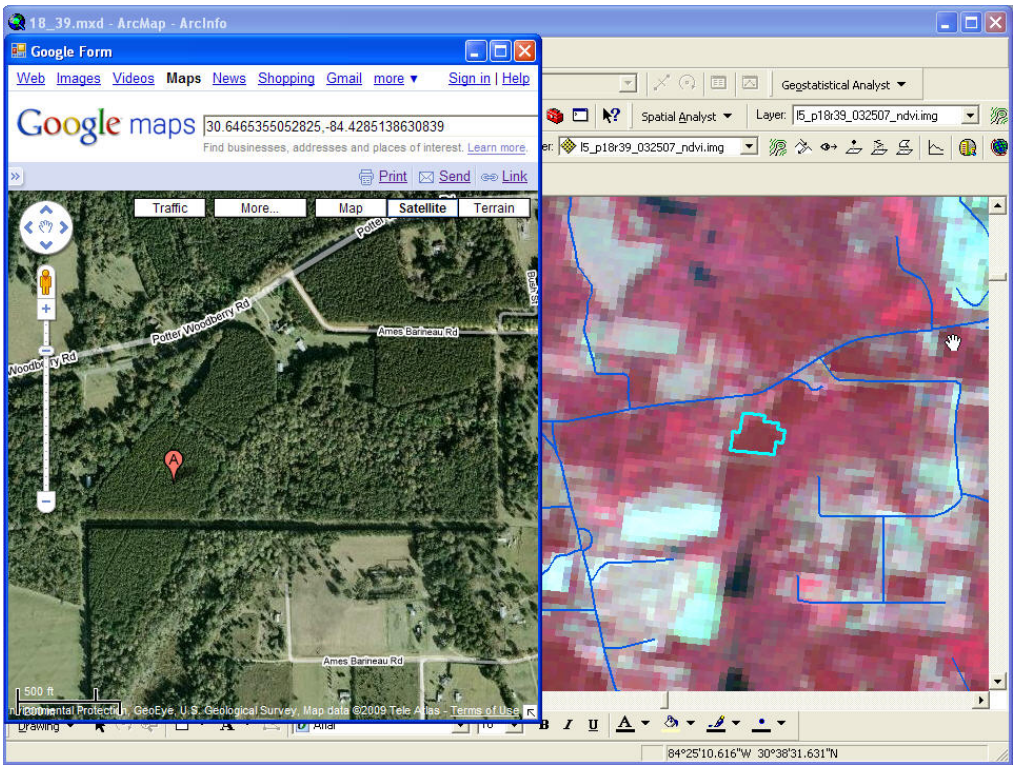


**Figure 4.** An ArcObjects tool to link polygons in ArcGIS to Google Map.

## Support Vector Machine Regression Model

The software package R has several implementations of the SVM such as e1071, kernlab, klaR, and svmpath (Karatzoglou, 2006); the package e1071 is the first implementation (Dimitriadou, 2005) package and is based on the famous LIBSVM library. The svm() function in e1071 provides an efficient way to train and predict SVM model along with visualization and parameter tuning methods.

The standard process of SVR modeling in R is straightforward, and this process includes loading data, SVR model selection (includes kernel and cost parameter selection), training model, and using the selected model for prediction. The key feature is to select most accurate kernel function and cost parameter. Package e1071 has implemented four kernel functions: linear, polynomial, sigmoid and radial functions. An R program is designed to

select the best kernel function and cost parameter. The cost parameter is the parameter used to penalize the misclassification, for example a cost parameter of 0.5 means that the routine will not over fit the data where as a cost parameter of 3 will tend to over fit the data. The method is to apply cost parameter from 0.01 to 3 with intervals of 0.01 and evaluate the prediction error for each of the four kernel function to train SVR. The graph that shows the lowest prediction error based on the sample dataset is the most suitable kernel for the analysis. Figure 5 shows that the radial kernel achieves the lowest error and therefore the highest accuracy with a minimum error of 26,786 at a cost parameter of 1.10. Therefore these were the parameters used in the subsequent analysis.

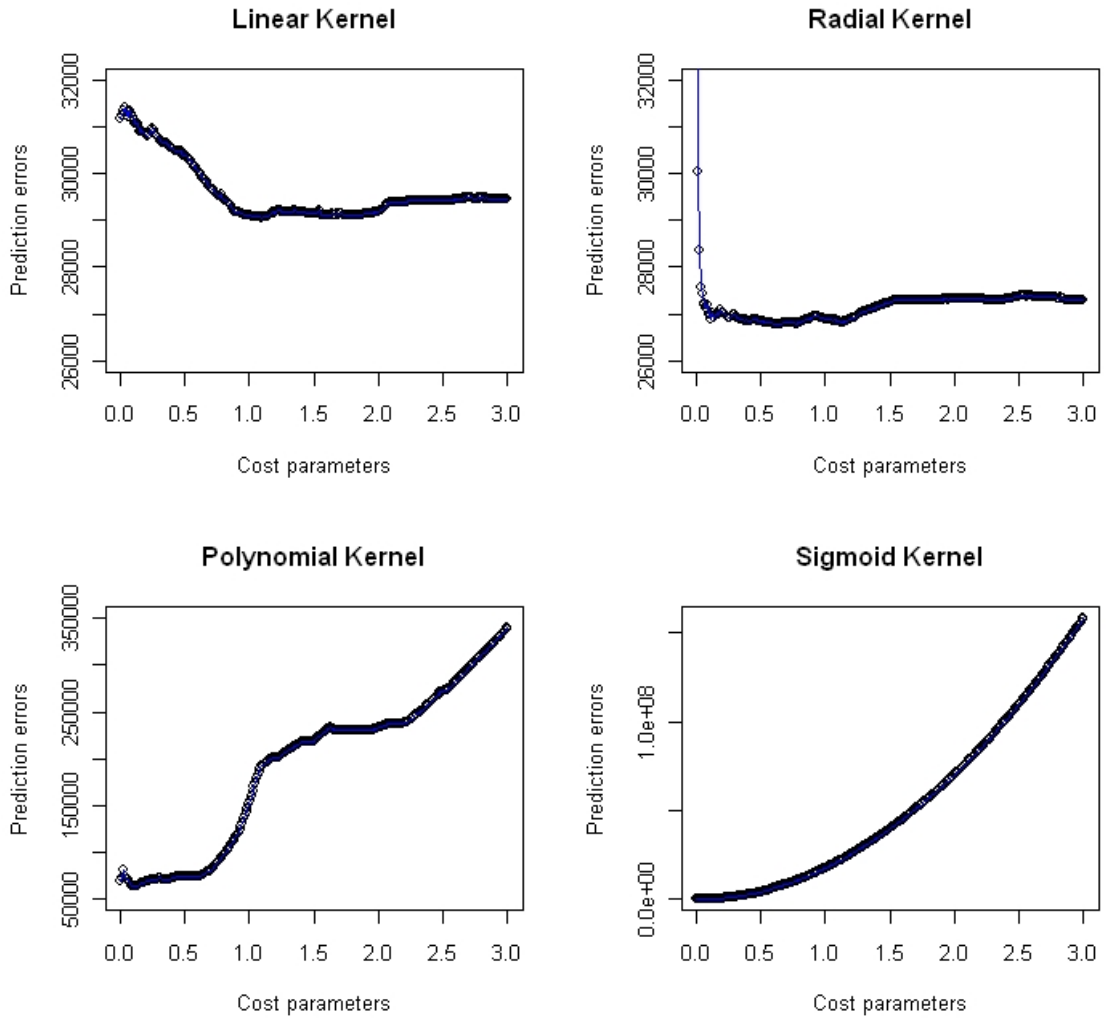


**Figure 5.** The relationship between prediction error and cost parameter for the four kernel types for the sample dataset.

## RESULTS AND DISCUSSION

As stated in the methods section, one fifth of the reference data was held to evaluate the SVR predicted value. Figure 6 is the scatter plot of SVR predicted value against reference data set. The R-square value is 0.6416 which is a relative high value considering that the spatial resolution of dataset is 30 m and that there are possible errors in the reference data collection, suggesting that this method may be appropriate for forest canopy cover extraction.

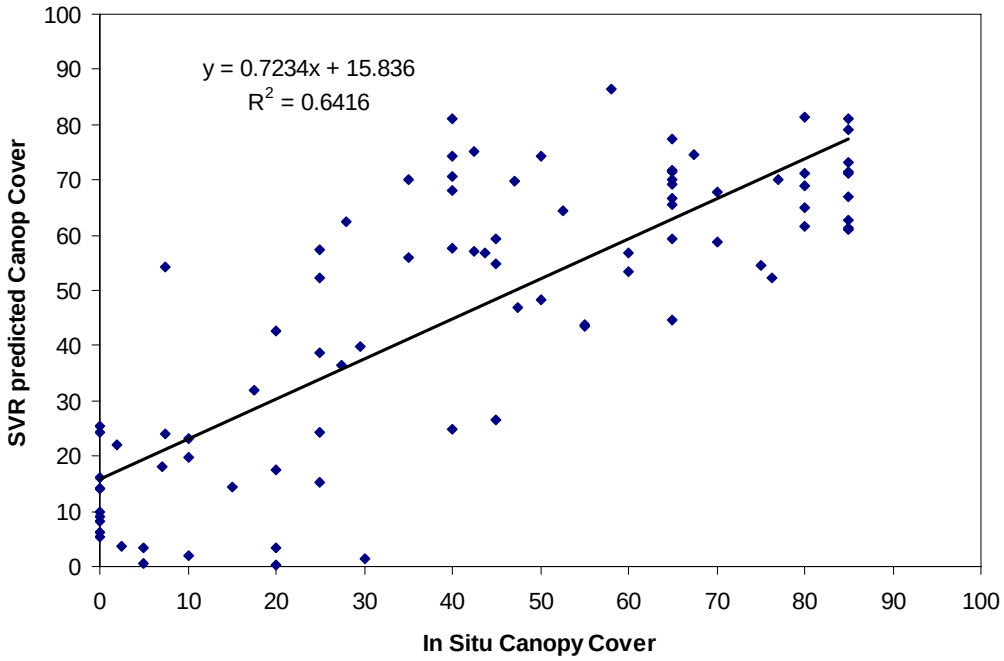


**Figure 6.** Scatter plot of predicted canopy cover vs. reference dataset.

In the above model (Figure 6), several spatial attributes were used. The spatial variables were area of polygon, rectangular fit of polygon, and ratio of length to width. An analysis was conducted to see whether the removal of these variables changed the SVR prediction accuracy. A new model was run to test whether the removal of spatial variables changed SVR prediction accuracy using exactly the same training data, test data and reference data. Figure 7 shows the scatter plot of predicted canopy cover without using spatial variables. The R-square value is 0.6554, which is a little higher than the previous model. This indicates that for this dataset spatial variables did not improve canopy cover predictions thus their inclusion did not increase model accuracy. This also shows that the inclusion of more variables in a SVR model will not necessarily increase SVR prediction accuracy.

## CONCLUSION

Forest canopy cover is one of the most important indicators of forest ecosystem. In this study, we have presented a new method to perform forest canopy cover regression based on support vector machines. An experiment was carried out in Northern Florida, and the test indicated that the SVR provided an efficient method to model forest canopy cover. The study also indicated that the inclusion of more attributes in the SVR model will not necessarily increase the accuracy of SVR model prediction, and this implies that a careful examination of sample data is necessary before applying all the attributes into SVR model.

This study also demonstrated a way to tune SVR model and to select most optimized kernel function and cost parameter. And this study indicated that radial basis kernel is the best kernel function for SVR model for this test data. Further study is needed to understand whether this is true for canopy cover for all Landsat imagery.

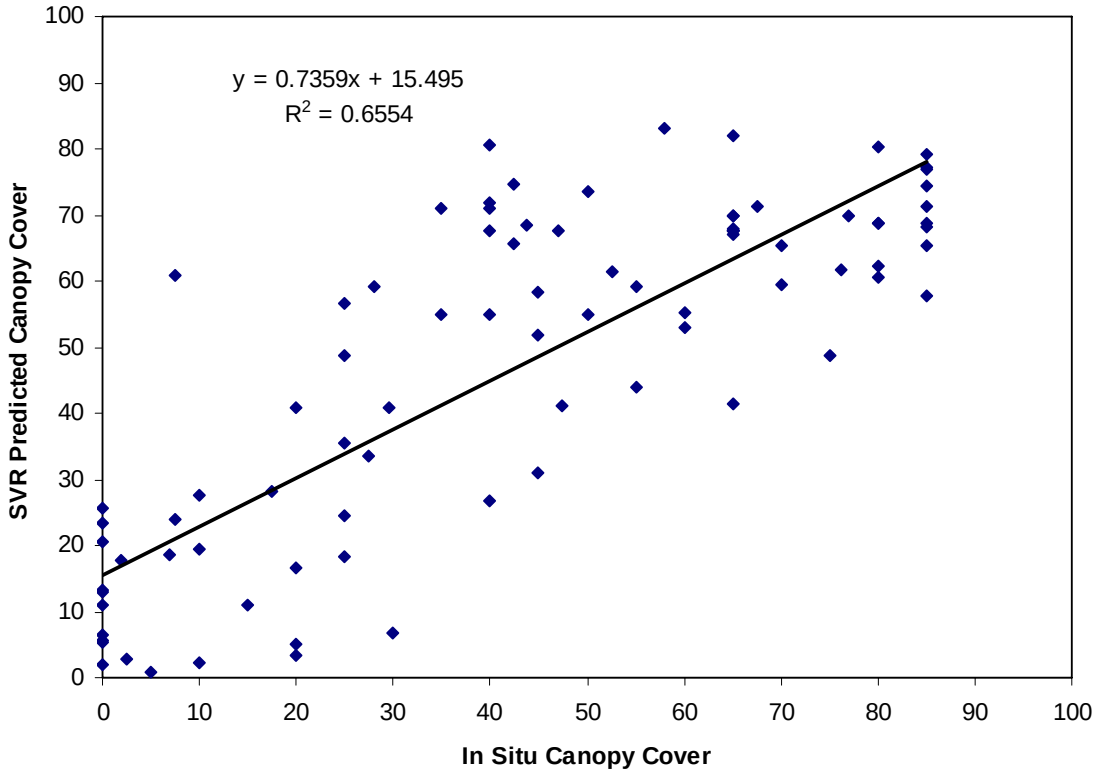


**Figure 7.** Scatter plot of SVR prediction without using spatial attributes.

## REFERENCES


Carreiras, J., José M.C. Pereira, and João S. Pereira, 2006. Estimation of tree canopy cover in evergreen oak woodlands using remote sensing, *Forest Ecology and Management*, 223(1-3): 45-53.

Dimitriadou, E., K. Hornik, F. Leisch, D. Meyer, and A. Weingessel, 2005. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien, Version 1.5-11, URL http://CRAN.R-project.org/.

Drucker, H., C.Burges, L. Kaufman, A. Smola, and V. Vapnik, 1996. Support vector regression machines, *Advances in Neural Information Processing Systems 9*, NIPS 1996, 155-161, MIT Press.

Gonzalez, Woods, and Eddins, *Digital Image Processing Using MATLAB*, Prentice Hall, 2004.

Huang, C., L. Yang, B. Wylie, and C. Homer, 2001. A strategy for estimating tree canopy density using Landsat 7 ETM+ and high resolution images over large areas, *Proceedings of the Third International Conference on Geospatial Information in Agriculture and Forestry*, in Denver, Colorado, 5 -7 November, 2001.

Jennings, S.B., N.D. Brown, and D. Sheil, 1999. Assessing forest canopies and understory illumination: Canopy closure, canopy cover and other measures, *Forestry*, 72: 59-74.

Karatzoglou, A., D. Meyer, and K. Hornik, 2006. Support vector machines in R, *Journal of Statistical Software*, 15(9):1-28.

Larsson, H., 2000. Linear regressions for canopy cover estimation in Acacia woodlands using Landsat-TM, -MSS and SPOT HRV XS data, *International Journal of Remote Sensing*, 14(11): 2129-2136.

Nemmour, H., and Y. Chibani, 2006. Multiple support vector machines for land cover change detection: An application for mapping urban extensions, *ISPRS Journal of Photogrammetry & Remote Sensing* 61:125–133.

Vapnik, V., 1998. *Statistical Learning Theory*, Wiley, New York.

Zhang, R., and J. Ma, 2008. An improved SVM method P-SVM for classification of remotely sensed data, *International Journal of Remote Sensing*, 29(20): 6029-6036.