SAMPLING INTENSITY AND NORMALIZATIONS: EXPLORING COST-DRIVING FACTORS IN NATIONWIDE MAPPING OF TREE CANOPY COVER

John Tipton, Gretchen Moisen, Paul Patterson, Thomas A. Jackson, and John Coulston

ABSTRACT

There are many factors that will determine the final cost of modeling and mapping tree canopy cover nationwide. For example, applying a normalization process to Landsat data used in the models is important in standardizing reflectance values among scenes and eliminating visual seams in the final map product. However, normalization at the national scale is expensive and logistically challenging, and its importance to model fit is unknown. Cost also increases with each location sampled, yet appropriate photo sampling intensity relative to the FIA grid has yet to be explored. In addition, cost is also affected by how intensively the photo plots themselves are sampled with a dot count, and the effect of reducing the number of dots on predictive models is also unknown. Using intensively sampled photo plot data in 5 pilot areas across the United States, we address these three cost factors by exploring the effect of a normalization process of Landsat TM data on model fits of tree canopy cover using Random Forests regression, the relationship between the sampling intensity of photo interpreted plots and model fit, and the relationship between the number of dots for each photo interpreted location and model fit.

INTRODUCTION

The National Land Cover Database (NLCD, http://www.mrlc.gov/) for 2011 will contain a map of tree canopy cover that will be a spatially explicit map-based data on percent tree canopy cover is used for forest management, estimates of timber production, determining the potential for and extent of fire danger and other management issues across the United States. The 2001 NLCD provides map-based estimates of percent tree canopy cover along with land cover and percent impervious cover (Homer and others 2004). The NLCD is a periodic product with an update cycle of approximately five years. However, because of funding constraints the percent tree canopy estimates were not updated for the 2006 NLCD. For the 2011 NLCD the U.S. Forest Service Forest Inventory and Analysis program (FIA) will take the lead on developing the percent tree canopy cover layer.

FIA is uniquely positioned to lead the development of the 2011 NLCD percent tree canopy cover layer. First, FIA uses a probabilistic sample design that covers all lands (forest and non-forest) and can be easily intensified for geospatial modeling purposes. Second, the FIA program is beginning to make percent tree canopy cover estimates for all sample locations. This provides an opportunity to leverage data collected as part of the FIA program to develop predictive models used to produce percent tree canopy map products. To this end, a pilot study was carried out in 2010. The pilot study was designed to answer specific research questions and estimate costs for developing the 2011 NLCD percent tree canopy cover map.

Creating a tree canopy cover product that encompasses the entire country presents many questions that must be answered before prototype or production mapping can begin. Consequently, a pilot project was launched that included five study areas, one each in Georgia, Michigan, Kansas, Oregon, and Utah. Within each study area, over two thousand photo plots were photo-interpreted by an interpreter looking at a grid overlaid on an aerial photo of each plot. At each of the 105 points on the grid, the interpreter determined if the point was a tree or not, and this response was used to calculate percent tree cover.

Using data from the pilot project, several issues are addressed in this paper to support production of mapping of tree canopy cover nationwide. First, the number of samples plays an important role in the quality of the model. It is important to find a balance between the quality of model fit and concerns of cost. Second, normalization of Landsat TM images is important because adjacent Landsat scenes on a map are not taken on the same day. Because of this, when a mosaic of multiple images is constructed, there will be

John Tipton, Colorado State University, Fort Collins, CO 80525
Gretchen G. Moisen, Research Forester, US Forest Service Rocky Mountain Research Station, Ogden, UT, 84401
Paul Patterson, Statistician, US Forest Service Rocky Mountain Research Station, Ogden, UT, 84401
Thomas A. Jackson, Colorado State University, Fort Collins, CO 80525
John W. Coulston, Research Forester, US Forest Service Southern Research Station, Knoxville, TN, 37919

seams in the image where the raw reflectance values for one image are not equal to the reflectance values of the adjoining image. Normalization of one image to another using the overlap between two images will remove the visual seam in a map, but the effect of normalization on how well a model predicts percent tree canopy cover has not been explored. Third, at each sample locations an estimate of percent tree canopy cover was made using a simple dot grid approach. The pilot study design used 105 dots however, if the same information can be obtained with fewer points, we can trim costs and maintain the quality of the model. Consequently in this paper, we explore the effects of sample size, normalization and number of dots on predictive models of tree canopy cover.

METHODS

Percent tree canopy cover data was collected for five study areas in the coterminous United States (Figure 2). The standard FIA sampling grid (1 plot per 2400 ha) was intensified fourfold to 1 plot per 600 ha using the techniques described by White et al. (1992). At each sample location a 105 point dot grid covering a 90m by 90m area was developed. At each of the 105 points, a photo interpreter determined if the point was a tree or not, by examining high resolution digital aerial photography collected in 2009 (USDA 2009). The percent tree canopy cover for each sample location was defined as the number of points intersecting tree crowns divided by 105 and was used as the dependent variable for random forest model development.

The independent variables came from a variety of sources but they were primarily Landsat 5 data and vegetation indices derived from Landsat data (e.g. normalized difference vegetation index, tasseled cap). Additionally, digital elevation models and derivatives (e.g. slope, aspect) were also used as potential independent variables for random forest model development. The Landsat data were available as normalized mosaics and non-normalized mosaics. Because each study area covered multiple Landsat scenes differences in spectral values among scenes arise because of differing collection data and atmospheric effects. The non-normalized data had no correction for these effects. The normalized data accounted for these effects by standardizing reflectance values from a target scene to a reference scene based on the overlap among scenes.

The specific modeling tool used was Random Forests, implemented in R using the library RandomForests (Liaw and Wiener 2002). Random Forests is a machine learning process that uses decision trees for classification and regression. The algorithm computes many trees, with each tree getting a "vote," with the final model being a majority decision (categorical variables) or average (continuous variables). For each node in those trees, a subset of explanatory variables is randomly selected and a dichotomous split in the data is made based on the largest decrease in the MSE of the data. To get the final model, the process is run for 500 trees, and the results are averaged. Each tree is constructed using a randomly selected set of the data where approximately one-third is held "out of bag" and can, therefore, be used as a validation data set and as a measure of model fit. Our measure of model fit is called pseudo R^2 and it represents the relative amount of variation in the data that is explained by the model. Pseudo R^2 is calculated as 1-MSE/Var(y) where the pseudo R^2 is calculated individually for every tree in the forest, then averaged over all trees to compute the final value.

To investigate the question related to sample size, we performed an iterative sampling process where, for each iteration, plots were randomly sampled from our study site, a model is fit using the RandomForest command, and the measure of model fit (pseudo R^2) is recorded. Then, for the next iteration, the number in the sample was increased by 20 plot locations and so on until the number in the iterative sample equaled the total sample size for the study site. When plotting the pseudo R-squared values against the number of study site samples, we applied a lowess smoothing curve for each of the study site locations to get a visual indication of the asymptotic behavior. From this method, we were able to get estimates of the variance of the fit of the model as well as to determine the asymptotic behavior of model fit relative to sample size.

The simulations described above were performed for both the data set that was normalized (corrected for differences in Landsat scenes) and for the data set that was not normalized. This allowed us to also explore the asymptotic behavior of the model fit relative to normalization.

The final question had to do with the number of dots used for the photo interpretation grid. For each study site we sampled 500, 1000, and 1500 study locations and calculated the percent tree cover based on randomly sampling a number of photo dots. We started with sampling one dot, and then fit a Random Forest model and recorded the pseudo R^2. The process was then iterated, increasing the number of dots by one each time. In the plot of model fit versus the number of dots, we applied a lowess smoothing curve to see patterns in the simulations and to get a visual indication of the asymptotic behavior relative to number of dots. Also, estimates of the number of man-hours needed to complete a prototype of the same size with different number of sample plots and numbers of dots were produced. This assumed 3 minutes for loading each sample plot picture and another 3 minutes to count all 105 dots.

RESULTS

From these simulations we were able to get an understanding of what intensity sampling intensity provides the most information for the least cost. In Figure 2 we call attention to the smoothed curve of pseudo R^2 versus the number of sample plots for the non-normalized data in Oregon. Looking at the spread of the simulated model fits we see that between 1000 and 2000 sample plots the variation in simulated pseudo R^2 drops off quickly. This is of interest because the default FIA sampling intensity grid for this study site is approximately 1500. A similar pattern is seen in Figure 3, in which the variation in simulated pseudo R2 drops off quickly between 1000 and 2000 sample plots for the other four study sites.

Figures 1 and 2 also show the effect of normalization on model fit. When looking at the plot of sample size versus pseudo R^2 for Oregon in Figure 2 we see that there is little difference in the fit of our model with regards to whether the data was normalized or not normalized. When looking at the four plots in Figure 3 we see the same pattern in Georgia, Utah, and Michigan, but we have different results in Kansas. In the Kansas plot we see that the normalized data model outperforms the normalized data model, but the difference is small (at 4000 sample points the difference in pseudo R^2 between the normalized and non-normalized models is about .03). These results indicated that normalization plays a very minor role in the quality of model fit, and we made the decision to consider only the non-normalized data set for the rest of the analyses.

In Figures 3 and 4 we are looking at the plots of pseudo R^2 versus number of dots on the photo grid. By looking at the plots of number of dots versus pseudo R^2 in Figure 4 we see that in Oregon we are not getting more information by including more than 40 dots. This is evidenced by the inflection in the lowess smoothing curve on the plot. The same pattern is repeated in Figure 5 for the other study sites. By combining the recommendations of using non-normalized data and roughly 1000 sample plots per study site we are able to make estimates of the amount of man-hours needed to complete a study site of similar size. Figure 6 shows the amount of person-hours needed versus the number of photo interpretation grid dots for 500, 1000, and 1500 sample plots. Using our assumptions that each image takes three minutes to load and three minutes to calculate tree cover using all 105 dots, we plotted the number of photo grid dots versus time for 500, 1000, and 1500 sample plots. From this we can see that if we used 1000 sample plots with 40 dots we would expect one person to finish all five study areas in about 12 weeks.

DISCUSSION

By looking at the smoothed curves for the non-normalized data in Figures 1 and 2 we see the relationship between the number of sample plots and the precision of the model fit as measured by pseudo R^2. We see that between 1000 and 2000 sample plots the variation in pseudo R^2 decreases rapidly versus the number of sample plots when compared to larger sample sizes. This suggests diminishing returns in model fit when increasing the number of sample plots beyond values in the 1000 to 2000 range. This suggests that we can get good model relative to cost in the 1000 to 2000 sample plot range, which also happens to be approximately the FIA standard sampling intensity grid for each study site.

Choosing to use only non-normalized data to fit a Random Forests model has major implications for the budget of the project. Normalization is an expensive and time consuming process, especially on a scale the size of the entire United States. Our results indicate that the Random Forests model performs equally well using either normalized or non-normalized data. From this result, we are able to make recommendations to get a higher quality product for less cost. However, the visual effects of not normalizing are still under investigation.

Because a human observer will be used to measure percent tree cover in the final product, using fewer dots will decrease the time the observer will spend on each photo, which will decrease the overall cost of the project. Since it appears that we gain little in terms of model fit when considering more than 40 dots, this suggests that we can reduce the person-hours needed for the prototype.

CONCLUSION

Because there are limited resources available it is important to get an understanding of the behavior of the sampling protocols and model fits relative to the costs of the process. The recommendations in this paper give guidelines for the next prototype phase of the NLCD Canopy Cover project.

ACKNOWLEDGEMENTS

We would like to thank everyone who has worked with this pilot project for collecting the data, taking the time to photo interpret the images, and provide counsel on this project. Also we would like to thank Dr. Jean Opsomer for his time and assistance.

LITERATURE CITED

Liaw, A.; Wiener, M. 2002. Classification and Regression by randomForest. R News 2(3), 18—22.

Breiman, L. 2001. "Random Forests." Machine Learning. 5-32.

Breiman, L.; Friedman, R. A.; Olshen, R. A.; Stone, C. G. 1984. Classification and regression trees. Pacific Grove, CA, USA: Wadsworth.

Homer, C.; Haung, C.; Yang, L. [and others]. 2004. Development of a 2001 national land-cover database for the United States. Photogrammetric Engineering and Remote Sensing. 70, 829-840.

U.S. Department of Agriculture. 2009. National Agriculture Imagery Program, Salt Lake City; U.S. Department of Agriculture, Farm Service Agency, Aerial Photography Field Office. Information available: http://www.apfo.usda.gov/FSA/apfoapp?area=home&subject=prog&topic=nai

White, D.; Kimerling, A.J.; Overton, W.S. 1992. Cartographic and geometric components of a global sampling design for environmental monitoring. Cartography and Geographic Information Systems 19: 5-22.

Figure 1—Location and extent of the five pilot study areas.

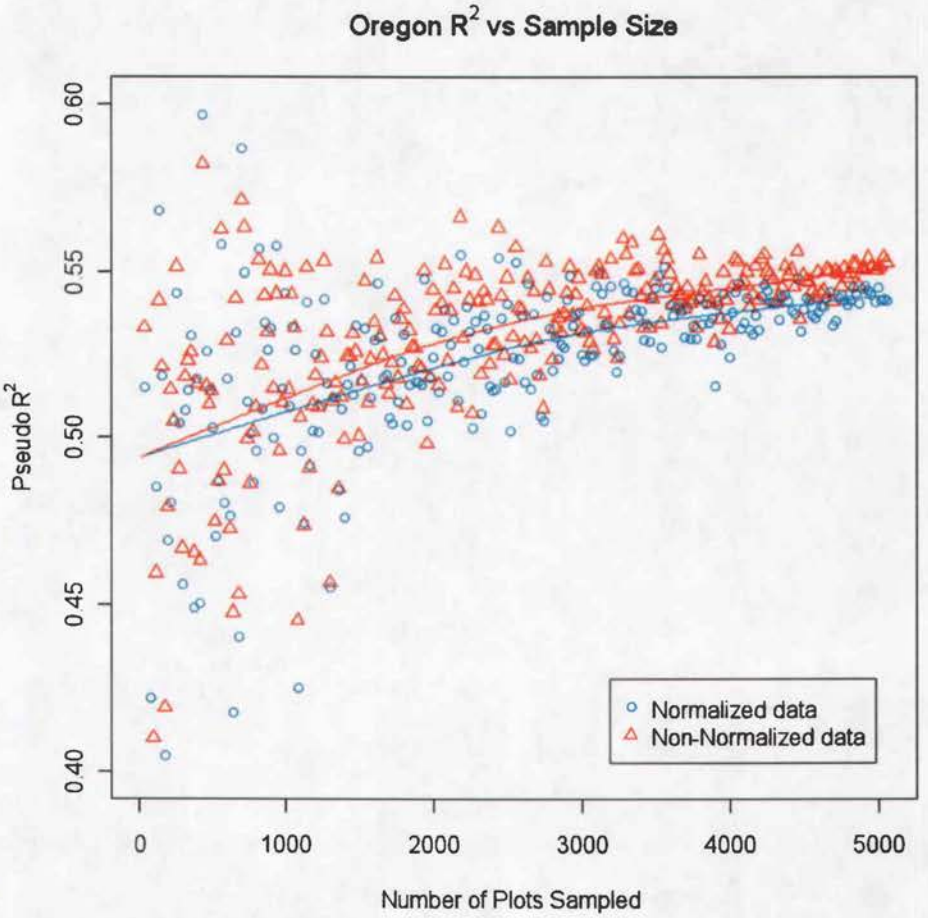

Figure 2—Shows the pseudo-R^2 values plotted against the number of plots sampled for Oregon for both the normalized and non-normalized data sets with the solid lines representing a lowess smoothing curve.

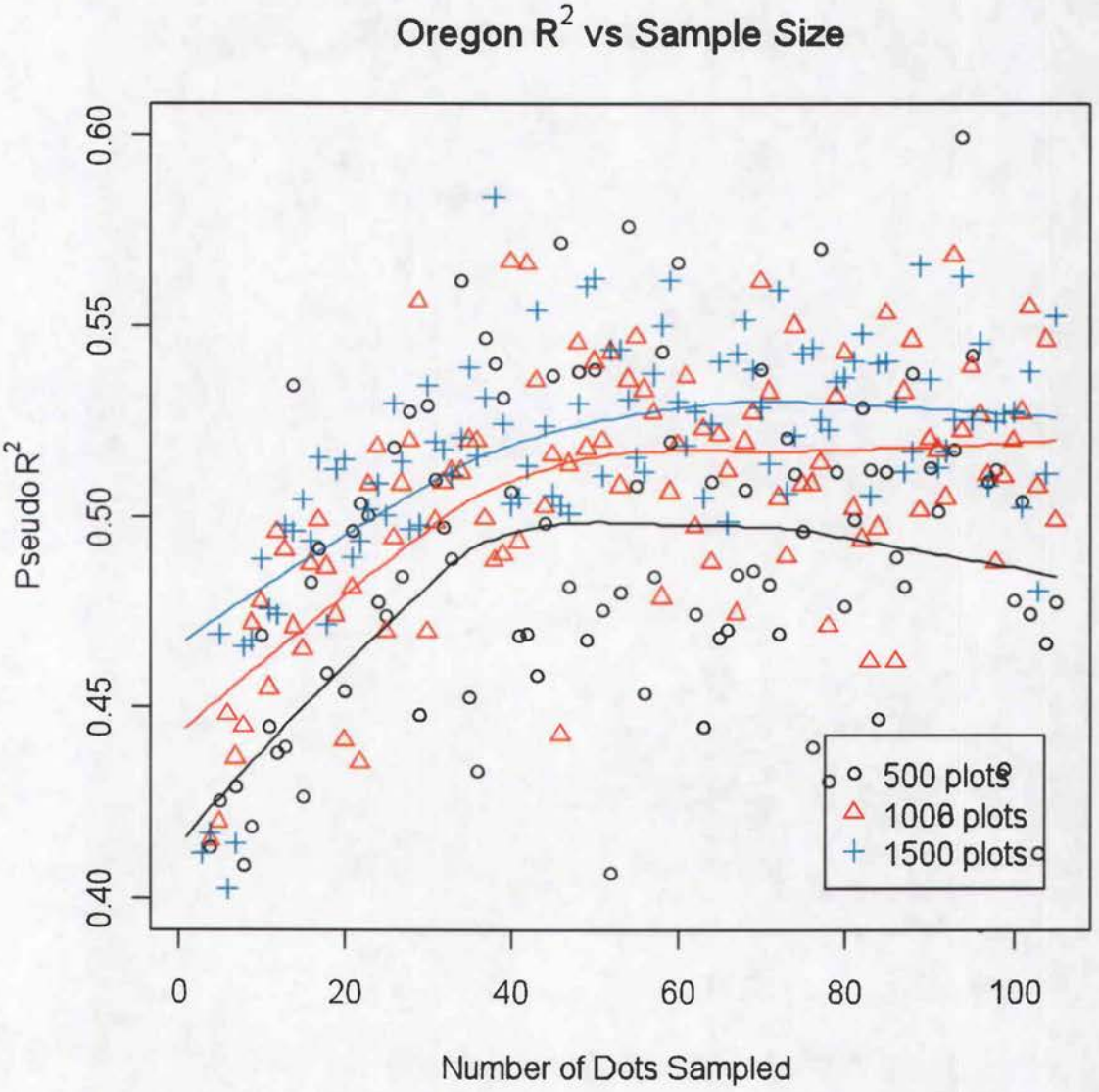

Figure 4—Shows the pseudo-R^2 values plotted against the number of dots sampled for Oregon, for both the 500, 1000, and 1500 sample plots with the solid lines representing a lowess smoothing curve.

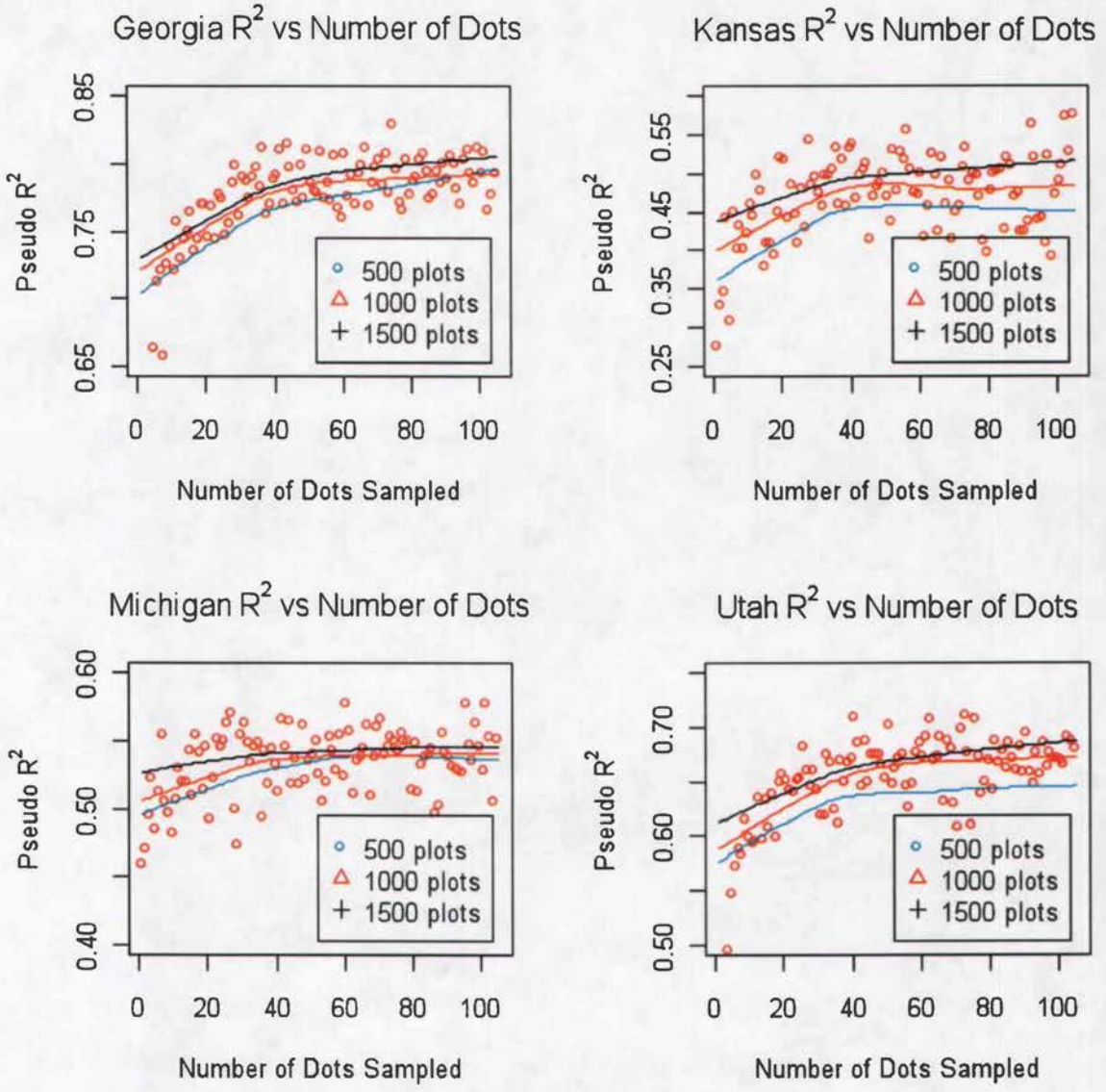

Figure 5—Shows the pseudo-R2 values plotted against the number of dots sampled for Georgia, Kansas, Michigan, and Utah, for both the 500, 1000, and 1500 sample plots with the solid lines representing a lowess smoothing curve.

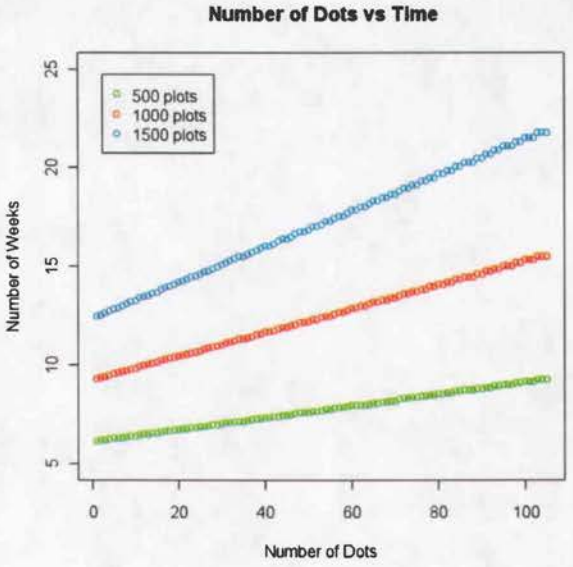

Figure 6—Shows the amount of time to complete a prototype of similar size toon the five study sites versus the number of dots used in photo interpretation for 500, 1000, and 1500 sample plots.