

RESAMPLING REMOTELY SENSED DATA TO IMPROVE NATIONAL MAPPING OF FOREST CONDITIONS

Raymond L. Czaplewski

USDA Forest Service, Rocky Mountain Research Station, Fort Collins, CO, USA

Abstract: The USDA Forest Service (FS) built a 250-m resolution (6.25-ha pixel) dataset for the entire USA, including Alaska. It assembles multi-seasonal multispectral MODIS, Landsat derivatives, and other national geospatial datasets. There are currently over 250 layers in this “stack” of full-coverage wall-to-wall geospatial data. A separate “modeling data set” includes training data from the FS Research and Development’s Forest Inventory and Analysis (FIA) program for approximately 160,000 forested plots, which are registered with the corresponding pixel in the “stack” of national geospatial data. Limitations include a scale-mismatch between the 1-ha field plot and the 6.25-ha remotely sensed pixel, field plots that straddle multiple pixels, and the legal requirement to protect confidentiality of the exact location of each plot.

One partial solution is resampling to a smaller pixel size, such as a 150- by 150-m (2.25-ha) pixel. This would diminish the scale-mismatch, while recognizing the random errors in measuring the exact position of each FIA field plot. However, a 1-ha FIA plot could straddle up to four contiguous 2.25-ha pixels. A complementary solution would be resampling a single 2.25-ha area centered on the best known position of each field plot. This could improve the empirical relationships between remotely sensed data and corresponding field measurements in the training dataset, which might improve statistical models that predict field measurements from remotely sensed data.

By Public Law, FIA must keep the exact locations of its field plots confidential. In a “stack” of 250 geospatial layers, most pixels will have a unique set of geospatial values among the millions of pixels in an image. This presents a risk that locations of FIA field plots could be deduced from the modeling dataset. However, it is unlikely that any of the 2.25-ha pixels on the full-coverage resampling grid would have exactly the same multivariate geospatial data, or “fingerprint”, as the one 2.25-ha pixel that is centered on the field plot. Supplemental swapping of geospatial data among similar pixels in the full-coverage “stack” would further obscure the location of training plots in the modeling dataset. This offers the potential to distribute training data for the 2.25-ha pixel “stack” to duly authorized external FIA agents without the risk of compromising confidentiality of the exact locations of field plots.

INTRODUCTION

The USDA Forest Service (FS) built a 250-m resolution (6.25-ha pixel) dataset, called the “stack”, for the entire USA. The “stack” refers to a set of 250 pixel-grids or images. It contains no training data from the FS Research and Development’s Forest Inventory and Analysis (FIA) program. However, a separate “modeling data set” was created by intersecting FIA plot centers, which links to FIA field measurements, with the “stack”

extracting corresponding full-coverage geospatial predictor variables at those points, and joining them in a flat file. The “stack” and modeling data sets are separate although related files. The unique value of these two datasets comes from integration of FIA field data from each of approximately 160,000 forested plots with their corresponding pixels in the “stack” of geospatial data. Each FIA field plot represents approximately 1-ha and is re-measured every 7 to 10 years at an approximate cost of \$2,000 per plot. After a labor-intensive data integration effort, FIA data are available to FIA agents to conveniently train empirical classification and regression models that use geospatial data as predictor variables, and field measurements as the response variables.

The FS has three short-term objectives mapping for the USA: (1) develop a forest biomass map; (2) develop a forest stand type map; and (3) develop a detailed tree-species composition map. These geospatial data would provide important inputs to National Forest risk mapping. Longer-term objectives might include continuous development of additional geospatial data and model predictions using a shared geospatial data infrastructure. All products recognize the random errors in predicting detailed field conditions at the pixel level with remotely sensed data.

LIMITATIONS OF THE EXISTING “STACK” AND MODELING DATASETS

The first limitation of the “stack” is the substantial scale-mismatch between the 1-ha FIA field data and the 6.25-ha geospatial pixel data; field measurements only characterize a small portion of a 6.25-ha pixel, and it is likely that the field data do not realistically represent the field conditions of many 6.25-ha pixels.

The second limitation is misregistration error caused by a resampling grid that is misaligned on the field plot locations. For example, a 1-ha field plot could straddle up to four contiguous 6.25-ha pixels. The current practice picks the single pixel from the full-coverage resampling grid that covers the largest portion of the field plot, even though part of the plot, occasionally the majority of the plot, could be outside of that pixel.

Halme and Tomppo (2001) found that the second limitation is aggravated by registration and rectification errors. The combined spatial error varies from 0 to 70 m, the average being 10 m. Statistical models that predict field conditions with remotely sensed data are extremely sensitive to registration errors in coordinates of field plots and rectification errors in matching remotely sensed pixels to their corresponding field coordinates. GPS can reduce registration errors, but not rectification errors in the GIS and remotely sensed data.

The third limitation comes from the need to keep the exact locations of FIA field plots confidential, which is legislated by the Food Security Act of 1985 (as amended by PL 106-113). Geospatial data for 250 themes in the “stack” and modeling datasets could, in principle, be used to exactly locate the 6.25-ha pixel containing the plurality of a FIA field plot because most pixels in a “stack” of 250 geospatial variables are likely unique. Therefore, public release of the FIA field data associated with the geospatial “stack” is

prohibited by Public Law. Therefore, the full capabilities of the geospatial “stack” are only available to duly authorized FIA agents.

POSSIBLE SOLUTIONS

One partial solution to the first limitation is resampling to a smaller pixel size. This paper suggests a 150-m (2.25-ha) pixel, although a similar case could be made for a 90- or 120-m pixel. Although not a completely satisfactory solution to the scale-mismatch, this pixel size could improve the match between the scales of field data and remotely sensed data, while considering the reality of misregistration caused by errors in measuring the exact position of each field plot (e.g., random GPS errors) and image rectification. For example, assume a positional error of 30-m in a random direction; the full 1-ha FIA field plot could exist anywhere within a 2-ha circular area. Data volumes with a 2.25-ha pixel increase modestly relative to a 6.25-ha MODIS pixel, but data volumes remain much smaller relative to a 0.1-ha Landsat pixel database.

One solution to the second limitation is to resample a 2.25-ha area centered on the best known position of each field plot (Figure 1). This area would rarely, if ever, be identical to the closest 2.25-ha pixel on the full-coverage resampling grid. This kind of resampling would be independently repeated for each field plot in the training dataset. Statistical prediction models would be fit with the geospatial predictor variables from 2.25-ha pixels centered on each field plot. The resulting prediction models would then be applied to the full-coverage resampled grid with 2.25-ha pixels.

1.A. Re-sampling frame for training data and model construction

1.B. Re-sampling frame for wall-to-wall mapping with models from step A.

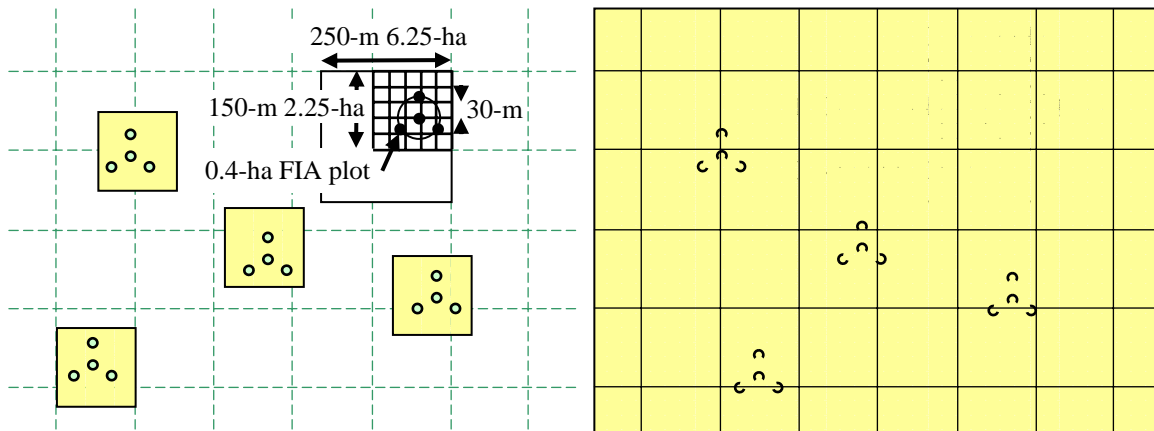


Figure 1. (A) Independently center a 150-m pixel on best known geospatial location of each FIA field plot. The purpose is to minimize effects of misregistration between remotely sensed and field data. The intentional misregistration of a training pixel with the full-coverage grid of pixels would make it very difficult to use the unique spectral “signature” of a training pixel in the modeling dataset to locate its geographic location within the full coverage “stack”. This could protect confidentiality of FIA plot locations, and it could allow FIA to publicly distribute a full geospatial database and associated FIA field data. (B) Apply models built in from training pixels in Step A to wall-to-wall coverage of geospatial data, acknowledging that no single pixel grid will ever simultaneously match well with field plots.

PLOT CONFIDENTIALITY IN A PUBLIC GEOSPATIAL DATABASE

The proposed solution to the second limitation (above) has a potentially beneficial consequence. It might allow the Forest Service to publicly offer its “stack” of remotely sensed data and corresponding field data in the modeling dataset to anyone, even those who are not FIA agents, without compromising plot location confidentiality. This is a partial solution to the third limitation.

It is unlikely that any of the four 2.25-ha pixels on the full-coverage resampling grid would have exactly the same multivariate geospatial data, or “fingerprint”, as the one 2.25-ha pixel that is centered on the field plot and that straddles those four pixels. The untested expectation is that the remotely sensed fingerprint of a field plot is unique within the immediate vicinity of that plot, and perhaps unique within the full “stack” of remotely sensed data. Use of tasseled cap and other linear transformations of spectral data (*e.g.*, principal components factors) would further complicate the decoding task if the coefficients in the transformation remain confidential; it would be difficult or impossible to deduce those transformations from the image data. Also, rounding real numbers to integer 8-bit values would add another obstacle to inferring locations of plot locations from the training dataset. If these expectations prove valid, unpopular methods such “fuzz-and-swap” and adding random noise to geospatial predictor variables are unnecessary to protect plot confidentiality.

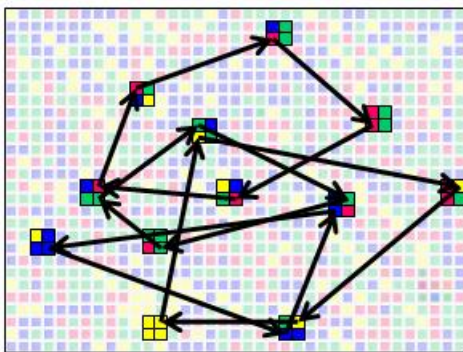
While this solution might sufficiently assure confidentiality of vast majority of FIA field plots, the solution might be insufficient for a FIA plot having a very rare spectral signature and when the FIA plot is centered within the 2.25-ha pixel in the full-coverage database. One solution would be to add some random noise to the spectral signature of rare training plots. However, this could also degrade statistical prediction models fit with those training plots. An alternative is to swap remotely sensed measurements of pixels near an FIA training plot within the full-coverage geospatial “stack” (Figure 2A.). Rather than affecting expensive training data, this solution would only affect a relatively small number of pixels in the geospatial “stack,” which includes millions of pixels. The idea is to encode pixel swapping in such a way that it would be almost impossible to decode the locations of training pixels within a full coverage geospatial “stack”.

For example, identify training pixels that have a rare signature (Figure 2A). For each such pixel, randomly pick one of the four possible 4x4 clusters of pixels that contain the rare training pixel. Randomly pick with replacement among these four pixels, assuring that the training pixel is selected at least once. Find spectrally similar pixels for each pixel so selected. Repeat the process a few times. Then, swap spectral signatures within each group of similar pixels. In order to deduce location of training plots within the geospatial “stack”, it would be necessary to first deduce which few of the millions of pixels have signature swapping.

The 150- by 150-m pixel used to form the geospatial “stack” would create more decoding challenges. Assume a particularly clever and persistent decoder could deduce which

pixels have swapped spectral signatures; the next challenge would be deducing which swapped pixel was actually a training pixel. Although no system can absolutely guarantee plot confidentiality, including the *status quo* in FIA, the proposed image stacking, resampling and pixel swapping should make it nearly impossible to identify training pixels in a full geospatial “stack”, even if the unique spectral signature of each training plot were known. This offers the potential to publicly distribute training data for the 2.25-ha pixel “stack” without compromising confidentiality of field data measured on private property. If correct, this would expand the customer base for FIA plot data and its geospatial “stack”, while improving confidentiality and transparency of geospatial analyses.

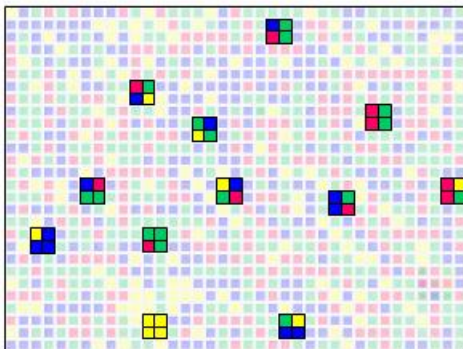
2.A Encoding: Form clusters of pixels associated with a training pixel, and swap geospatial data



2.B Decoding challenge: which of the millions of pixels have swapped geospatial data?



2.C Decoding challenge: which pixels with swapped geospatial data similar to a training pixel?



2.D Decoding challenge: Which pixel is the one training pixel collocated with an FIA field plot?

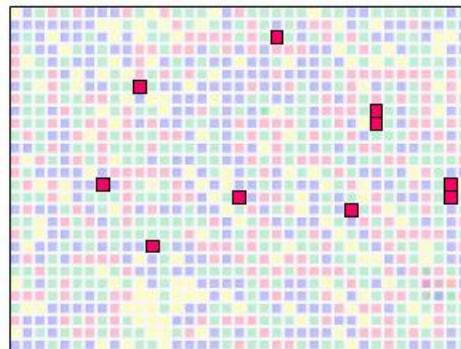


Figure 2 (A) Swapping spectral values for similar training pixels would make it unlikely to deduce which of the millions of 150- by 150-m (2.25-ha) pixels corresponds to one of the training pixels having a rare spectral “fingerprint.” This is an example of swapping associated with one training pixel. First, it would be a daunting task to deduce which clusters of four pixels have swapped spectral values. (B) The decoding challenge is made difficult by separate resampling of each training pixel and the full-coverage geospatial “stack”, and keeping confidential the parameters used to resample 0.1-ha Landsat pixels into 2.25-ha pixels. Use of linear combinations of the spectral data, such as the tassle cap transformation, and 8-bit integer round-off, would impose additional challenges. (C) Assuming that swapped pixel clusters could be deduced, which type of pixel is associated with a training pixel? (D) Assuming that this information could be deduced, which pixel is actually a training pixel? The hypothesis is that swapping pixel values would make it essentially impossible to deduce the exact locations of training pixels. Swapping geospatial data for a relatively few number of pixels does not degrade the empirical relationships between field measurements and geospatial predictor variables in the training dataset.

DISCUSSION

One assumption made here is that misregistration between field measurements and pixel data degrades the predictive power of statistical models, which are needed to produce maps of forest attributes from remotely sensed predictor variables. This assumption remains untested, and the gains in statistical modeling might not be substantial. Before applying the concepts presented in this paper, trials should be conducted on a diversity of pilot areas.

Building a modeling data set that resamples the 150-m pixel centered on each plot might be infeasible because of the computational burden. The computational process might be streamlined by subsampling small subsets of the original geospatial data inputs near nominal locations of FIA field plots and conversion of pixel data into vector format. Also, the positional inaccuracy of many older FIA plots can be up to 200-m, in which case subsampling might not improve the registration of field data to a 150-m pixel.

The resampling described in this paper assumes a linear combination (*e.g.*, a weighted average) of remotely sensed values from neighboring pixels in the immediate vicinity of a FIA field plot. There are several popular resampling options in image processing software systems that would satisfy this assumption, including cubic convolution and bilinear interpolation. In my opinion, the latter is preferable in the context of this application, but resampling options should be investigated in the pilot tests.

CONCLUSION

Creation of a national geospatial “stack” and modeling dataset puts the FIA program in the unique position to rapidly produce and update nationwide maps of detailed forest attributes. The FIA program has a credible history of quantifying the information quality of such maps, which might well serve the requirements under the federal Data Quality Act.

The potential for protecting plot confidentiality would facilitate distribution of the “stack” and modeling datasets to duly authorize external agents of the FIA program, and ease concerns over filling requests to intersect FIA plots with agents’ own stacks of geospatial data. However, this is a policy decision that can only be made by the FIA program managers.

ACKNOWLEDGMENTS

I thank the following reviewers for their helpful suggestions on an earlier draft: Kenneth Brewer, William Cooke, Mark Finco, Ronald McRoberts, Gretchen Moisen, Stephen Prisley. However, I remain solely responsible for the content of this paper.

CITATIONS

Halme, M. and Tomppo, E. 2001. Improving the accuracy of multisource forest inventory estimates by reducing plot location error - A multicriteria approach. *Remote Sensing of Environment* 78(3):321-327.