Preliminary Notes on NPS and LANDFIRE Comparison

Author: Steve Stehman
Initial: June 29, 2011
Revision 1: August 22, 2011

Task from LANDFIRE [Jim Smith, Don Ohlen contacts]

Compare the NPS assessment results with LANDFIRE assessment results to develop an overall, integrated review of LANDFIRE spatial product quality. The purpose is to inform potential users of LANDFIRE EVT—help them understand and apply the agreement assessment results in their applications.

Introduction

The notes that follow focus almost exclusively on the existing vegetation map (EVT) provided by LANDFIRE and accuracy issues associated with EVT. A relatively large number of Applications Projects have been completed, but these appear to be case study applications comparing LF to other local products in terms of very specific fire-related applications that could use LF data. Because I am much less familiar with the products that are incorporated in the Application Projects, I will not attempt to include those in these initial notes.

1. NPS and LANDFIRE (LF) assessments of EVT

1A. What are the differences between the approaches, and why? (Questions in bold specified in the work proposal.)

LANDFIRE (LF) Approach

a. Broadly regional in scope (i.e., sample data available for nearly national coverage)

b. Validation sample selected from the LF plot database (validation sample is a "holdout" sample of plots not used for developing the LF classification)

c. Existing plot data in the LF database were obtained from a variety of sources, and not all sources implemented a probability sampling design to select sample locations

d. The sample of "holdout" plots is a probability sample, but the original full sample of LF plots in the LF database is not a probability sample, so a probability sample from a non-probability sample database is still a non-probability sample and any biases or problems of representation that exist in the LF plot database will be present in the validation sample of holdout plots

e. An EVT assessment could be based only on those plots that do have a probability sampling origin, for example FIA plots; such an analysis has not yet been done for LF, but would be a useful complement to the EVT assessment using all holdout sample plots

f. The Landfire validation sampling design was strongly influenced by several constraints: 1) no additional sample plots obtained exclusively for the purpose of validation which limited the assessment sample to existing plots in the Landfire plot database; 2) keep the sample for

validation separate from the sample plots used to train the classifier(s); 3) sample plots representing rare classes were not common so extracting a large number of these plots for validation was not done to keep these plots available for training the classification.

g. Two problems resulting from the decisions in f) above are that the validation sample is not a probability sample and rare classes are not well represented. A partial remedy to the first problem would be to revise the analysis so that the estimates adjust for any deviations of the sample from known features of the EVT map (e.g., if certain classes are clearly over-represented in the sample, the analysis can be adjusted to compensate for that mis-representation). It is not possible to fully eliminate problems of a non-probability sample.

NPS Approach

a. More local in geographical coverage (limited to a few National Parks). Specifically, we have documents for the following assessments of Landfire EVT:

1) Glacier
2) Grand Teton
3) Delaware Water Gap
4) Rocky Mountain
5) Zion

These five parks were assessed with the same (or a very similar) protocol. The sampling design was stratified by NPS vegetation classes so the samples include a better representation of rare NPS classes, and assuming some EVT rare classes are associated with the rare NPS classes, some rare Landfire EVT classes will have larger sample sizes than would be the case in the Landfire holdout sample. The report for Rocky Mountain and Zion (by David Eckhardt) provides details on the analysis that indicate the stratified design was properly taken into account in the analysis. The methodology used in these five assessments is sound based on information available in the documentation.

6) Aniakchak National Monument

7) Denali

The two Alaska assessment were based on a variety of sources of reference sample data, some new sample sites and some existing sites. The sampling design has some of the same issues of representation that is present for the Landfire holdout sample, but these two assessments still provide useful information regarding EVT quality.

8) Hawaii

9) Onslow Bight (NC)

These two assessments are based on overlaying GAP maps with Landfire EVT.

A variety of other "Application Projects" have assessed other Landfire outputs, for example, FRCC. These assessments are not discussed here.

b. Probability sample of areas within parks using a stratified sampling design, with the strata defined by NPS mapped classes (i.e., classes present for NPS vegetation maps). As stated in the previous comment, the NPS design is entirely defensible. The fact that the NPS strata are based on NPS classes is not a design flaw – it simply means that a potentially better stratification exists for sampling EVT classes, but the results obtained with an "imperfect" stratification are not biased if the data are analyzed properly and the details presented in the NPS documentation indicate the correct analysis was done.

c. Some NPS plots were omitted from the analysis of LF EVT because it was not possible to assign a LF class based on the available NPS plot data and going through the sequence table process of labeling.

d. LF approach has an advantage of covering a much broader geographic area, but suffers from the problem that it is a sample of existing plots in the LF plot database; the NPS sample is a probability sample and therefore has a stronger statistical basis for inference (generalization), but area included in the NPS studies is a much more limited geographic representation.

1B. Are the results different, with a distinction?

a. Comparing the results of the LF and NPS assessments will be difficult because of the differences in the regions covered by the assessments. We will likely not have comparable data in the sense of an NPS sample and an adequately large LF sample covering a common region because the LF holdout sample is too sparse within a National Park. The suggestion was raised during the August 19 telecon of whether we could look at all sample plots within a National Park as a common plot basis for a comparison. This could be done. We know that there is the danger that by including some plots in the validation sample that were used for training we will get an optimistic assessment of accuracy, it may still be informative to examine accuracy based on all available plots so that the sample size is larger.

b. A reason the two results could differ would be if the LF plot database badly represents the general population. This could occur if the LF plots are heavily weighted toward difficult or easier classes so that the accuracy results are correspondingly too pessimistic or too optimistic. It will be possible to compare the class proportions in the LF sample to the LF proportions in the area evaluated to examine if the sample is representative of the area mapped. That is, the distribution of LF sample points to EVT classes should approximate the relative proportions of the EVT classes in the EVT map. If the proportions are not correct, it is possible to re-analyze the data to make an adjustment to proportional representation.

c. The LF approach is not stratified so it is not designed to evaluate rare EVT classes (i.e., the lack of stratification will lead to small sample sizes for rare classes and imprecise estimates of accuracy of these classes). The NPS design is stratified, so accuracy estimates for rare classes may be reasonably precise. These differences between NPS and LF designs may result in the samples looking different, but the accuracy estimates, when properly weighted for the stratified NPS design, would be comparable even though two different sampling designs were implemented. Estimates from two different probability sampling designs of the same population would be expected to be the same (taking into account sampling variation attributable to each design).

1C. What are the commonalities in the approaches, and the results?

a. Both approaches use the Landfire sequence table to assign a reference class so in that sense they are similar in response design.

b. As stated in note c) within1B above, the accuracy results would be expected to be similar if the NPS and LF approaches were applied to a broad area. The primary difficulty will be that the NPS and LF results will be representative of different regions.

1D. Are the differences important?

a. Because we will lack a good way to compare quantitatively the LF and NPS assessments, it will be difficult to make a statement about the importance of differences.

b. The starting premise is that any differences should be minor and potentially attributable either to natural variation in the estimation process (with the standard errors of the accuracy estimates providing a measure of this variation) or differences in the regions assessed.

1E. Do they tell a similar story, or very different?

If the story is different, it will be necessary to identify the potential causes of the differences. The obvious starting points would be to examine the sampling designs (e.g., the existing plot database of LF is not a representative sample, a design was implemented improperly) and analysis protocols to look for problems. Potential sources of differences between NPS and Landfire sample estimates:

1) Differences of geographic coverage
2) Different sampling designs. Even if NPS and Landfire assessments covered the same region, we would expect different sampling designs to produce samples that have different results.
3) Sample to sample variation (i.e., even if NPS and Landfire used the same sampling design in the same region, two samples still yield different results – the standard error quantifies this variation)
4) Protocols for labeling reference data are differ (although both use sequence table, perhaps there are differences in how the input data are obtained)

1F. Can we say anything about products downstream from EVT?

The main use of the accuracy assessments for downstream products would likely be via uncertainty or sensitivity analyses. For example, how much does the downstream product output change if EVT data were different? The classification error rates found in the validation studies can inform the sensitivity analyses by providing ranges for how the EVT product might change. In general, these types of sensitivity analyses are rarely conducted. What the current Landfire validation does provide is a regional (by map zone) description of accuracy of the more common classes. The regional description is important because it gives some sense of geographic variation in the results. The usual error matrix approach also provides the user some indication

of what types of misclassifications are occurring, and the user can decide which misclassifications are problematic to the downstream use of the EVT data. Often accuracy results are reported for one or more "aggregated" legends in which related classes are combined into one class. This repackaging of the data may be helpful information for some applications.

2. Identify and document other LANDFIRE assessments that have been performed, and include those in the discussion if possible and appropriate.
The assessments we have "in hand" were documented under 1A.

3. What are the uses of these quality assessments—what is their value, and to whom?
a. The primary use (initially) of the quality assessments will be as documentation – how accurate is the EVT product? This information can also be used for quality control purposes in that users and producers may judge that the quality is not acceptable and that a revision is suggested. The assessments may be regionalized to some extent, but only to the broader geographic regions envisioned for primary applications of LF

b. The results from the broader-scale, holdout sample assessment (e.g. LF) may not be adequate (in terms of sample size) to report accuracy by small regions (even as small as a National Park).

c. We can approach question #3 by considering the following two groups: 1) Value and uses of accuracy information for National Leaders (Directors, Congressional, etc.); 2) Value and uses for Regional and local managers (National Park, BLM District, National Forest, Fish and Wildlife Refuge, etc.)

d. A common problem with accuracy assessment is that users want information at a regional level that is smaller than what accuracy assessment sample sizes typically can support (for large-area mapping projects). For any subregion of a map, if accuracy results are desired for that subregion the subregion in effect becomes its own sampling problem. The sample size needed to provide precise estimates of accuracy within that subregion are nearly the same as the sample size required to provide precise accuracy estimates nationally. So any single user interested in a single subregion will likely not be satisfied with the sample data available from a map zone or national assessment. Some level of geographic generalization will need to be invoked by users to gain an understanding of the accuracy for their subregion of interest.

e. Providing the overall and class-specific LF EVT accuracy values may lead to a knee-jerk condemnation of LF because of low accuracy for the full EVT set of classes. Regional and local managers may be more attuned to accuracy issues of large-area vegetation and land-cover mapping and may understand the value of the EVT data even if accuracy is relatively low. The accuracy assessment results provide broad measures of EVT data quality along with information on what classes are typically misclassified, but it is also necessary to understand how these errors propagate through applications based on EVT and to get some idea of whether the errors have a substantial practical impact.

4. List some examples or case studies of how each of the above groups could use them and their potential value.

It is a good question to ask what users actually do with information from an accuracy assessment. It is common for users to use reported accuracy when discussing inputs to their application. But it would not be an easy task, for example, to take an error matrix, and use that information to propagate the classification error through the application of the map. The error propagation analysis would probably be as difficult or more difficult to construct than the application that is using the EVT product. This might be a task for investigation by someone with expertise in error propagation (e.g., Jim's colleague Steve Prisley).

5. What would you recommend in the future? What can we do as a Program to improve this assessment process, or make the results more useful to the various user communities?
a. Will there be future or further assessments of LF EVT? The high cost of collecting data will be the major deterrent to assessment of LF EVT. This leaves two primary options: 1) using existing reference data (e.g., data in the LF plot database as was the case for the initial holdout sample assessment, or attempting to use data such as NLCD reference land-cover data, which may be difficult to relate to EVT classes); and 2) Collaborative data collection for validation (e.g., if NPS is validating their own vegetation maps and their plot data can be run through the LF classification system, then the NPS data may become a primary source of independent LF validation data; FIA is the other major source of potential data). At the moment, my recommendations would be: 1) LF continues to use the holdout sample approach based on the LF plot database; 2) consider using NPS reference data as a good source for localized, detailed evaluation of LF EVT (even if NPS doesn't actually use the LF map); 3) Explore what can be done in terms of an assessment based strictly on FIA plot data.

b. To enhance utility of accuracy assessment: provide printed recommended guidance for users of LF EVT on how to interpret or use the EVT accuracy assessment results. The interpretation guidelines would inform the reader of the definitions of the basic accuracy estimates (overall, user's and producer's accuracies, or commission and omission error) and their interpretation. For example, the different interpretations associated with omission error and commission error could be reviewed. Some information related to the application of area estimation should be mentioned. This is the "non site specific" accuracy of the product (i.e., how well do the mapped areas correspond to the reference areas of the classes).

c. It might also be possible to obtain comparisons of LF EVT with other maps. This may be a low priority activity given the associated difficulties of such comparisons (e.g., different legends, spatial resolution, extent, etc.).

6. Deliverables
a. Document discussing the LF and NPS assessments.

b. Publish the LF EVT holdout sample methodology and results. The Discussion section of the manuscript could address issues related to NPS assessments and future options for validating LF EVT.

7. Other Issues/Thoughts from Telecon of August 19, 2011.

a. Current focus will be on what to do with the existing Landfire and NPS assessments.
b. If we start with the notion that both the NPS and Landfire assessments are valid but will give different results, where does that leave us? If we further agree that an NPS map for any specific park created by NPS will be better than Landfire for that park (Landfire was not tailored to any particular park), what issues does that leave us to resolve regarding the NPS assessment? If we brought in the NPS folks for a discussion, what would the agenda items be?
c. I'll raise the question (and this was something we discussed very briefly at the end of the PQWT's previous existence) that if the primary interest is in products downstream from EVT, is there a way to assess directly those downstream products? The main drawback would be that the basis of validation (direct assessment) is the existence of a "truth" against which the classified product can be compared, and sometimes it isn't feasible to obtain that truth.
d. The point was raised about the educational opportunity created by the Landfire assessment. How can we create informed users of accuracy assessment results related to applications of the product? What can the accuracy data tell us and what cannot be gleaned from the data? This might not require much explanation. For example, accuracy assessment results can provide generalized information about which classes are mapped accurately and what types of classification errors are present (very spatially explicit results), and it can tell us how well the area of a class is captured by the map (this is called "non site specific accuracy" in the jargon). The information is general in the sense that it does not tell you if your backyard is classified correctly – it only tells you what the likelihood is that your backyard is classified correctly. The accuracy assessment results do not tell you directly what impact classification error in the product will have on your application of the data. The accuracy results may give you helpful data to assess the impact on your application, but it will require additional analysis.
e. Because the Landfire national product is being revised, the specific validation results from the Landfire holdout sample are not of great interest because the product has been superseded.

8. Next Steps (tentative pending further discussion)
a. We can do a variety of re-analyses of the Landfire validation data and comparisons with the NPS results, but we need a clear purpose for doing so. Along these same lines, we could add comparisons of Landfire validation to results from validations of GAP products. Before we do that, we need to establish what objectives these analyses would accomplish.
b. Develop the idea of how to use the Landfire validation as an "educational opportunity". There could be technical issues associated with the Landfire methodology (which would feed into the objective of how we would design a future assessment) and also issues related to how users use the results of an accuracy assessment.