# Characterization and visualization of the accuracy of FIA's CONUS-wide tree species datasets

*Rachel Riemann*[1] *and Barry T. Wilson*[2]

[1] USDA Forest Service, Northern Research Station, Troy, NY. [2] USDA Forest Service, Northern Research Station, St. Paul, MN.

## Abstract

*Modeled geospatial datasets have been created for 325 tree species across the contiguous United States (CONUS). Effective application of all geospatial datasets depends on their accuracy. Dataset error can be systematic (bias) or unsystematic (scatter), and their magnitude can vary by region and scale. Each of these characteristics affects the locations, scales, uses, and questions to which a dataset is best applied, and the risk involved in doing so. This study uses a suite of assessment metrics to characterize the type, magnitude, frequency, and spatial location of errors in this large dataset. Results are examined with respect to tree species growth habits, level of stand dominance, spatial-distribution characteristics, and the number of plots on which it occur and for any persistent local errors occurring throughout the datasets.*

**Keywords**: accuracy characterization, continuous variables, comparative assessment, species distributions, FIA, visualization.

## 1. Introduction

Tree species abundance and distribution have been modeled for the contiguous United States (CONUS) using an approach that integrates vegetation phenology derived from MODIS imagery, raster data describing relevant environmental parameters, and extensive field plot data of tree species basal area (ba), using the modeling techniques of k-nearest neighbors and canonical correspondence analysis (Wilson *et al*. 2012). In this approach, model predictions are calculated using a weighting of the 2$^{nd}$ through 8$^{th}$ nearest neighbors based on proximity in a feature space derived from the model, ensuring that any co-located plot is never used in the imputation of a corresponding pixel. The effectiveness of these geospatial datasets for applications such as ecosystem research, policy analysis, and planning is impacted by their accuracy.

Errors in modeled geospatial datasets can take the form of truncated distributions, loss of variability, and/or overestimation or underestimation of values – characteristics which can affect the utility of the dataset for particular applications. The error will be some combination of both random (unsystematic) error and bias (systematic error), and these errors can vary by subpopulation, by region, and by scale. Such inaccuracies do not necessarily render a modeled dataset useless, but they do affect the locations, scales, uses, and questions to which it can be applied, and the level of risk associated with it. Thus an effective assessment of such geospatial datasets requires a characterization of each of these facets of error.

The suite of assessment metrics described in Riemann *et al*. (2010) characterizes the comparative accuracy of continuous variables by providing information on the type,

magnitude, frequency, and location of errors in each dataset. The United States Forest Service (USFS) Forest Inventory and Analysis (FIA) database of field plot data is an available reference dataset of sufficient sampling intensity to take full advantage of these assessments at multiple scales. Together these metrics provide information that map consumers can use to gauge the appropriateness of the overall map products for various uses at different scales.

## 2. Methods

Datasets of live basal area ($m^2$/ha) for all 325 tree species inventoried by FIA across CONUS were developed following the methodology described in Wilson *et al*. (2012) and assessed for accuracy. Modeled results were compared to data collected on FIA plots for all states except Colorado, New Mexico, and Oklahoma where assessment data were not available at the time. A suite of assessment metrics was applied using the protocol described in Riemann et al. (2010). Agreement between data distributions was quantified using the Kolmogorov-Smirnov statistic (KS) (e.g. Feller, 1948). Systematic and unsystematic agreement were captured independently using the non-parametric Agreement Coefficient (AC) developed by Ji and Gallo (2006). Root mean square error (RMSE) was calculated to provide information on the magnitude of difference in data units ($m^2$/ha). All metrics were calculated at four different scales defined by tessellation of the area into a hexagonal mesh of various spacing of hexagon centers and area sizes: a 54,000 ha area (25km spacing, ~22 plots per hexagon), 216,500 ha area (50km spacing, ~89 plots per hexagon), 866,000 ha area (100km spacing, ~357 plots per hexagon), and 3.5 million ha area (200km spacing, ~1441 plots per hexagon).

In addition, hexagon-level metrics were calculated at each scale to characterize the spatial distribution of differences, and mapped to provide a visualization of both uncertainty and disagreement. To describe the uncertainty in our comparison dataset, confidence intervals (CI) for the $90^{th}$, $95^{th}$, and $99^{th}$ percentiles were calculated for each hexagon based on the field plots within the hexagon. To describe level of disagreement with respect to plot-based uncertainty, we also identified at each hexagon where the model-based mean fell with respect to these confidence intervals. Seventeen species were chosen for more detailed examination (table 1). The set included several from each region, both locally- and broadly-distributed species, and across a range of prevalence (from 0.27 to 5.4 percent of CONUS plots) and a range of typical stand dominance (median values from 2.5 to 100 percent of the total basal area in stands where it occurs). The set also included understory species (e.g. dogwood), species with very fragmented distributions (e.g. river birch), both open and closed canopy forest types, and species occurring on a variety of terrain.

## 3. Results

Metrics summarizing comparative assessment results over the entire study area serve to characterize the overall relationship between the modeled and plot-derived estimates for each species. Table 2 presents a number of metrics for the 17 species, however results for all metrics for all 325 species are available at: (http://www.fs.usda.gov/rds/archive/Product/RDS-2013-0013/. Across all scales, systematic agreement ($AC_{sys}$) is generally very high (> 0.9) except for those species that occur on the fewest number of plots (< 0.5%), indicating very little bias in the modeled

datasets. Examining the relationship of the modeled mean to the plot-derived 90th CI, no observable spatial pattern in disagreement is present at any of the 3 scales, as is illustrated for ponderosa pine in Figure 1. In Figure 2, this agreement at the 50k hexagon scale is presented alongside a map of the species distribution, and a scatterplot of plot-derived vs. model-derived 50k-hexagon level means for five of the species.

**Table 1:** The 17 species examined in more detail and some of their associated characteristics.

| Species name | U.S. region(s) | %plots in which spp occurs | avg ba where it occurs ($m^2$/ha) | median % of total stand ba where it occurs | Additional characteristics |
|---|---|---|---|---|---|
| loblolly pine (*Pinus taeda*) | southeast | 5.4 | 1.2 | 52.2 | widespread |
| black cherry (*Prunus serotina*) | northeast+ | 5.3 | 1.2 | 5.0 | widespread, locally concentrated |
| northern red oak (*Quercus rubra*) | east | 4.4 | 1.0 | 10.9 | widespread |
| sugar maple (*Acer saccharum*) | northeast+ | 4.3 | 1.0 | 14.2 | widespread, locally concentrated |
| Douglas-fir (*Pseudotsuga menziesii*) | northwest, Rocky mtns | 3.4 | 0.8 | 44.5 | widespread, locally concentrated |
| white ash (*Fraxinus americana*) | east, mid-west | 3.4 | 0.8 | 6.7 | of interest--pest risk |
| dogwood (*Cornus florida*) | east | 3.1 | 0.7 | 2.5 | understory species |
| ponderosa pine (*Pinus ponderosa*) | west | 2.8 | 0.6 | 46.8 | locally concentrated |
| quaking aspen (*Populus tremuloides*) | NE, midwest, W | 2.2 | 0.5 | 20.6 | widespread |
| Utah juniper (*Juniperus osteosperma*) | southwest | 1.9 | 0.4 | 76.5 | open canopy, locally concentrated |
| lodgepole pine (*Pinus contorta*) | NW, Rocky mtns | 1.7 | 0.4 | 34.8 | open canopy |
| Englemann spruce | Rocky mtns | 1.3 | 0.3 | 23.3 | relatively widespread |
| honey mesquite (*Prosopis glandulosa*) | TX, AZ, NM | 1 | 0.2 | 100.0 | open canopy, local |
| Rocky mtn juniper (*Juniperus scopulorum*) | west | 0.6 | 0.1 | 19.1 | open canopy |
| river birch (*Betula nigra*) | east | 0.3 | 0.1 | 8.3 | sparse distribution |
| ailanthus (*Ailanthus altissima*) | mid-east | 0.3 | 0.1 | 5.2 | of interest--invasive |
| bigleaf maple (*Acer macrophyllum*) | west coast | 0.2 | 0.0 | 6.7 | very local distribution |

**Table 2:** Four of the study-area-wide agreement metrics: systematic agreement ($AC_{sys}$, max = 1), unsystematic agreement ($AC_{uns}$, max = 1), root mean squared error (RMSE, $m^2$/ha), and the proportion of the plot-based mean represented by the 90th CI (propCI). Eastern and western U.S. are calculated separately, and one species has sufficient presence in both regions to require two entries.

| Species name | %plots in which it occurs (E or W) | 25k | | | | 50k | | | | 100k | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ACsys | ACuns | RMSE | propCI | ACsys | ACuns | RMSE | propCI | ACsys | ACuns | RMSE | propCI |
| loblolly pine | 10.1 | 1.00 | 0.88 | 0.80 | 0.65 | 1.00 | 0.96 | 0.44 | 0.49 | 1.00 | 0.99 | 0.23 | 0.39 |
| black cherry | 10.1 | 0.97 | 0.64 | 0.24 | 0.91 | 0.99 | 0.86 | 0.14 | 0.60 | 1.00 | 0.94 | 0.09 | 0.37 |
| northern red oak | 8.3 | 0.97 | 0.65 | 0.33 | 0.91 | 0.99 | 0.88 | 0.18 | 0.65 | 0.99 | 0.96 | 0.10 | 0.45 |
| sugar maple | 9.2 | 0.99 | 0.83 | 0.38 | 0.81 | 0.99 | 0.93 | 0.21 | 0.61 | 1.00 | 0.98 | 0.11 | 0.45 |
| white ash | 6.4 | 0.95 | 0.50 | 0.21 | 0.98 | 0.99 | 0.85 | 0.10 | 0.73 | 1.00 | 0.95 | 0.06 | 0.53 |
| Douglas-fir | 8.9 | 1.00 | 0.89 | 1.17 | 0.75 | 1.00 | 0.96 | 0.62 | 0.59 | 1.00 | 0.97 | 0.40 | 0.49 |
| dogwood | 4.5 | 0.97 | 0.48 | 0.06 | 0.99 | 1.00 | 0.83 | 0.03 | 0.72 | 1.00 | 0.96 | 0.01 | 0.51 |
| ponderosa pine | 5.6 | 1.00 | 0.87 | 0.55 | 0.87 | 1.00 | 0.93 | 0.36 | 0.67 | 1.00 | 0.98 | 0.18 | 0.57 |
| quaking aspen-east | 5.2 | 1.00 | 0.84 | 0.21 | 0.80 | 1.00 | 0.94 | 0.11 | 0.60 | 1.00 | 0.98 | 0.06 | 0.48 |
| quaking aspen-west | 1.8 | 0.96 | 0.70 | 0.39 | 1.07 | 0.98 | 0.91 | 0.19 | 0.94 | 0.99 | 0.94 | 0.14 | 0.85 |
| Utah juniper | 3.5 | 1.00 | 0.80 | 0.80 | 0.93 | 1.00 | 0.93 | 0.40 | 0.68 | 1.00 | 0.97 | 0.26 | 0.52 |
| lodgepole pine | 3.5 | 0.98 | 0.73 | 0.77 | 0.88 | 0.99 | 0.89 | 0.45 | 0.69 | 1.00 | 0.95 | 0.28 | 0.53 |
| Englemann spruce | 2.7 | 0.98 | 0.75 | 0.56 | 0.93 | 1.00 | 0.89 | 0.31 | 0.74 | 1.00 | 0.88 | 0.30 | 0.59 |
| honey mesquite | 2.1 | 0.94 | 0.75 | 0.27 | 0.88 | 0.98 | 0.91 | 0.15 | 0.67 | 0.99 | 0.96 | 0.10 | 0.49 |
| Rocky mtn juniper | 1.2 | 0.92 | 0.14 | 0.20 | 1.12 | 0.95 | 0.56 | 0.13 | 0.94 | 0.99 | 0.86 | 0.05 | 0.79 |
| ailanthus | 0.5 | 0.63 | 0.00 | 0.04 | 1.19 | 0.90 | 0.44 | 0.02 | 1.06 | 0.98 | 0.82 | 0.01 | 0.89 |
| river birch | 0.5 | 0.38 | 0.00 | 0.06 | 1.24 | 0.85 | 0.23 | 0.03 | 1.10 | 0.95 | 0.67 | 0.01 | 0.91 |
| bigleaf maple | 0.7 | 0.81 | 0.13 | 0.16 | 1.01 | 0.92 | 0.66 | 0.09 | 0.80 | 0.99 | 0.96 | 0.01 | 0.76 |

In general, the more plots on which the species occurs, the more data the model had to work with and the better the agreement, however there are also numerous exceptions. For example, a more concentrated distribution appears to decrease uncertainty (propCI) and improve agreement (e.g. honey mesquite vs. Rocky mountain juniper or Douglas-fir vs. white ash). Factors potentially influencing the unexpectedly low $AC_{uns}$ values at the 25k hexagon level for dogwood and white ash may be the very low proportion of the total stand ba/acre each tends to occupy (2.5% and 6.7%, respectively) or dogwood's presence as primarily an understory species, in addition to their relatively scattered spatial distributions. Results for all 325 species reveal numerous similar anomalies, such as redwood (*Sequoia sempervirens*) whose concentrated distribution, high basal area and stand proportion, and strong relationship to environmental factors likely contribute to its unexpectedly high levels of agreement given the low number of plots on which it occurs.

Hexagon-level maps reveal substantial spatial variation in agreement across the dataset. This agreement is frequently high (within the 90th CI) in areas where the species is more dominant and may be of most interest. Assessment reveals that the modeled datasets tend to estimate species occurrence out beyond the edges of inventoried species ranges, due in part to the assessment dataset missing information in nonforest areas, as well as the lack of field information in these areas (Wilson et al. 2012). Many of these areas can be identified by their extremely low ba/ha estimates. Similarly, of those hexagons where plot mean > 0, a greater proportion of those modeled estimates falling outside the 90th CI were above the 90th CI (red) rather than below it (blue), indicating an overall tendency of the modeled estimate to be higher than the plot-based estimate, in part for the same reasons described above.

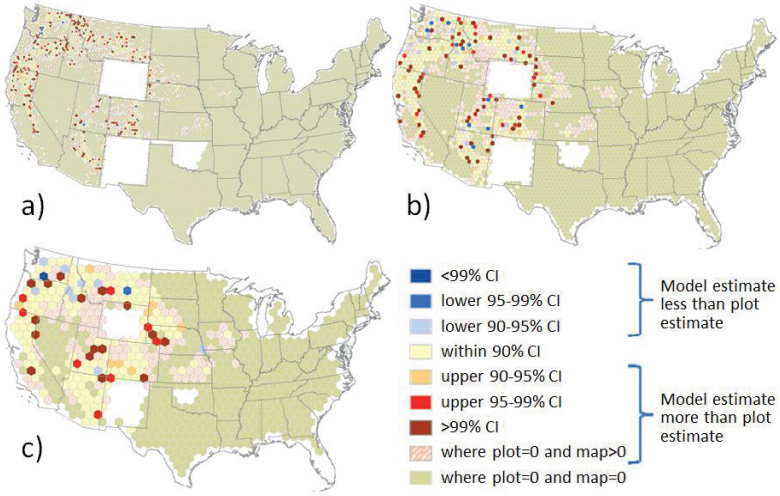


**Figure 1:** Relationship between mean modeled value and the plot-derived 90th confidence interval for ponderosa pine at the (a) 25k, (b) 50k, and (c) 100k scales.

## 4. Conclusions

Assessment of geospatial datasets is critical to their effective use and the metrics presented provide valuable additional information toward understanding and using these species datasets. Species characteristics, such as a concentrated distribution, tendency to occupy a high proportion of total stand basal area, occurring on a relatively large number of plots, status as an over- vs. understory species, or having a strong relationship to site characteristics or associate species, appear to affect the level of agreement. As the combination of these effects could be difficult to interpret, and/or predicting them would require more knowledge of an individual tree species than the average user can be expected to have at their disposal, the assessment metrics available with the datasets provide users with a direct look at dataset accuracy even without this additional knowledge.

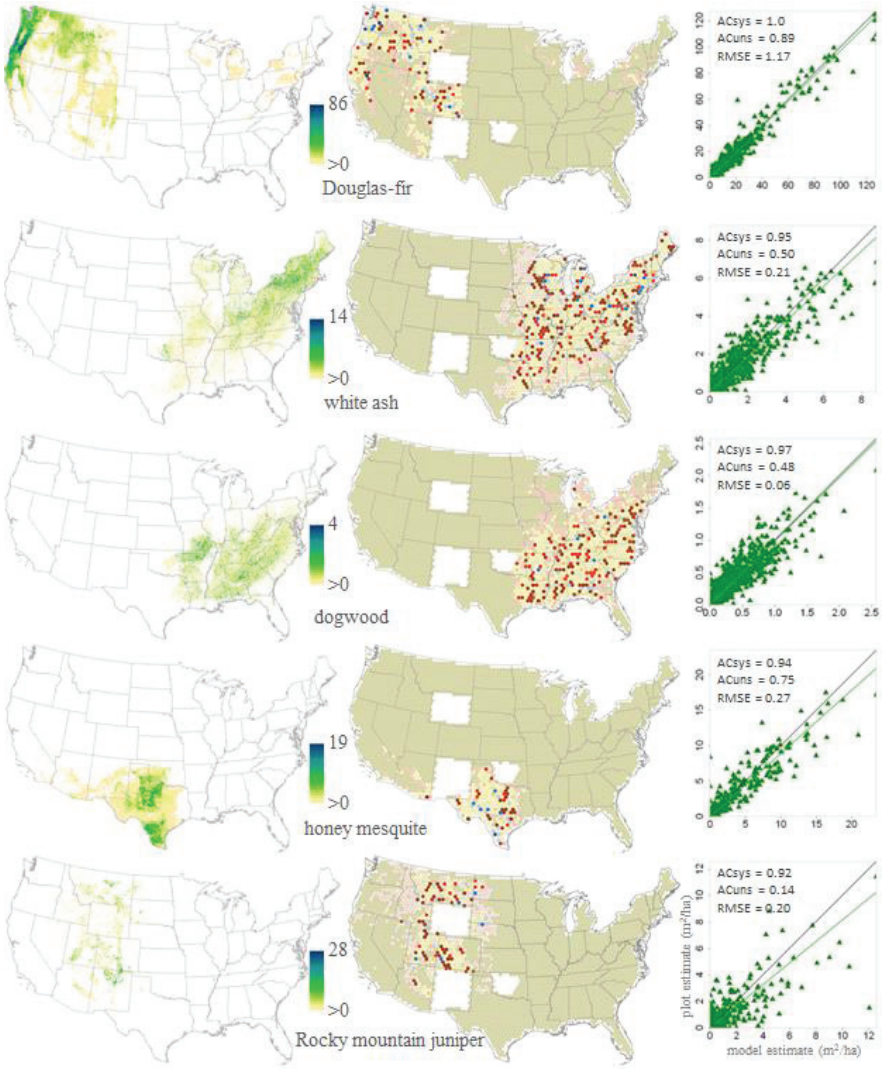


**Figure 2:** Maps of species basal area (sq. m/ha), 50k hexagon agreement map (legend in Figure 1), and 50k hexagon agreement scatterplot with both actual agreement (green line) and perfect agreement (black line).

Agreement coefficients calculated for the datasets as a whole provide a good indication of the overall level of systematic disagreement (bias) or unsystematic disagreement (scatter/noise) present. However even species with strong agreement coefficients can have local areas of high disagreement, and species with low agreement coefficients can have local areas of low disagreement. Maps of local area (hexagon) differences between modeled vs. plot-based estimates provide a valuable visualization of the spatial locations of these differences, which occur predominantly in areas where a species is less common or dominant.

Maps of local area (hexagon) differences that take into account the plot-based sampling error can provide a more realistic picture of where the map may be in error, as large differences in areas of high sampling error indicate that much of the disagreement

could be due to uncertainty in the plot-derived estimates as much as due to inaccuracies in the modeled dataset.

These species basal area datasets can be used with some level of confidence. Local uncertainties can be large (and propCI values high), but systematic agreement is very good, at all scales assessed, for the majority of species that occur on at least 0.5% of field plots.

## References

Feller, W. (1948), "On the Kolmogorov–Smirnov limit theorems for empirical Distributions", *Annals of Mathematical Statistics*, Vol 19(2):177–189.

Ji, L., Gallo, K. (2006), "An agreement coefficient for image comparison". *Photogrammetric Engineering and Remote Sensing, Vol.* 72:823–833.

Riemann, R., Wilson, B.T., Lister, A., Parks, S. (2010), "An effective assessment protocol for continuous geospatial datasets of forest characteristics using USFS Forest Inventory and Analysis (FIA) data". *Remote Sensing of the Environment,* Vol. 114:2337–2352.

Wilson, B.T., Lister, A.J., Riemann, R.I. (2012), "A nearest-neighbor imputation approach to mapping tree species over large areas using forest inventory plots and moderate resolution raster data", *Forest Ecology and Management*, Vol. 271:182-198.