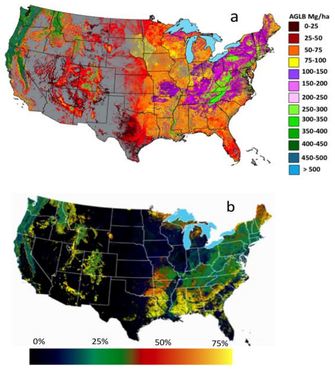
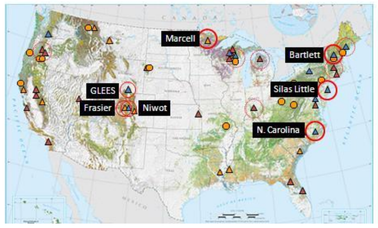


Decision Support and Robust Estimation of Uncertainty in Carbon Stocks and Fluxes

Stephen C. Hagen[1], Sassan Saatchi[2], Nancy Harris[3], Christopher W. Woodall[4], William A. Salas[1], Sangram Ganguly[5]

[1]Applied Geosolutions, Durham, NH
[2]Jet Propulsion Laboratory, Pasadena, CA
[3]Winrock International, Ecosystem Services, Arlington, VA
[4]US Forest Service, Northern Research Station, St. Paul, MN
[5]NASA Ames Research Center, Moffett Field, CA

Contact Stephen Hagen at shagen@appliedgeosolutions.com or 603-659-3363. **Work funded by NASA Grant NNX12AN72G**

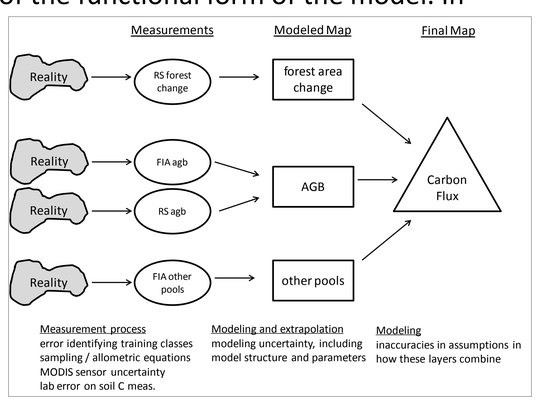


Introduction

The primary goal of our project (NASA Carbon Monitoring System – Saatchi PI) is to create detailed maps of forest carbon stocks and stock changes across the US to assist with national GHG inventories and thereby support decisions associated with land management. A comprehensive and accurate assessment of uncertainty in the forest carbon stock and stock change products is critical for understanding the quantitative limits of the products and for ensuring their usefulness to the broader community. However, a rigorous estimate of uncertainty at the pixel level is challenging to produce for complex products generated from multiple sources of input data and models. Here, we put forth a roadmap for assessing uncertainty associated with the forest carbon products provided as part of this project, which are generated by combining several sources of measurements and models.

Sources of Uncertainty

All of the ground measurements, remotely sensed observations, and process-based and statistical models are imperfect no matter how carefully obtained, managed, or processed. For instance, we will never know the true values of large forest population parameters, so we measure a subsample of all of the individuals. The inaccuracy of estimation of these parameters from a sample is termed the sampling error. The size of sampling errors can be controlled by the survey design and the size of the sample. After models are fit to data, substantial noise (i.e. residual errors) will certainly remain. This residual noise is due to both measurement and model uncertainty (i.e. noisy data and imperfect models), with model uncertainty potentially due to both parameterization and choice of the functional form of the model. In addition, the variance of these residuals can be heteroscedastic (i.e. not constant with respect to one or more of the independent variables). To calculate the error budget of total survey error, one needs to systematically identify and define all potential sources of error.



Framework for Robust Uncertainty Estimation

The bootstrapping approach to uncertainty assessment, also known as resampling with replacement, is more appropriate than conventional analytic methods for data with heteroscedastic errors. This method assumes that the observed data represent only one possible realization out of many, and reconstructs a large number of alternate realizations based on random resampling of the residuals. Bootstrapping brackets the range of unobserved values conditioned on the assumption of the model and its associated likelihood function. We estimate statistical uncertainty bounds associated with the final forest carbon stock and change estimates using a Monte Carlo-style sampling technique. With bootstrapping, we avoid making incorrect assumptions about the distribution of the underlying data sets, while combining the individual uncertainties from different sources.

Traditional analytic approach

Common analytic approach to combining uncertainty is to "sum by quadrature" . Given uncertainties from multiple sources (i.e. $\sigma_{allometric}$, $\sigma_{remote\text{-}sensing}$), total uncertainty is calculated by:

$$\sigma_{total} = \sqrt{\sigma^2_{allometric} + \sigma^2_{remote\ sensing}}$$

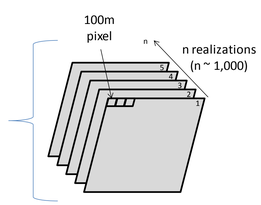
This approach typically assumes a Gaussian distribution of errors, which is often violated.

Uncertainty and Spatial Aggregation

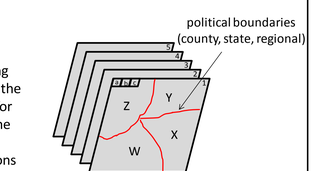
Create maps of:
Forest cover (2000, 2005, 2010)
Above Ground Biomass (2000, 2005)
Other carbon pools (2000, 2005)

Carbon flux (2000-2005, 2005-2010)

Each map will actually be a stack of n realizations of that variable. **Each pixel has n estimates, which form a probability density function (pdf) or distribution for that pixel.**

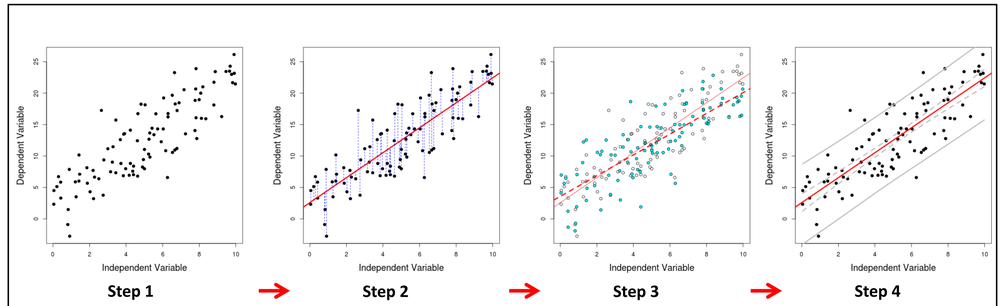


Spatial aggregation:
We preserve the stack of n underlying maps. As the final step, we first sum the pixels within the political boundary for each realization. Then we move to the next realization. Each political unit then has a distribution of n realizations from which summary statistics can be derived.



The value for county Z in realization 1 is a + b + c + ...

Monte Carlo approach using the bootstrap



Our bootstrapping approach (i.e. resampling with replacement) is similar to the approach many others have implemented (e.g. Robert and Casella, 1999; Press et al., 2003). Our recent applications include Hagen et al. 2006 & Harris et al. 2012. This approach provides a means of handling non-linear models and data with complicating characteristics such as heteroscedasticity. Above is an example using a linear ordinary least squares example. First, the model is fit to the original data (Step 1) and the residuals from this model are calculated (Step 2). Then, new realizations of the data are created (Step 3, cyan points) and a new model is fit to the new data (Step 3, dotted red line). This process is repeated n times (where n is large; > 1,000). The predictions from each iteration of the model are stored, as are the predictions plus a random residual. The predictions are sorted and the 2.5 and 97.5 percentiles are extracted to estimate the 95% confidence intervals (Step 4; grey dotted line). The predictions + random residual are also sorted and the percentiles are extracted to estimate the 95% prediction limits (Step 4; grey line).

Why do we use an MC framework rather than quadrature to propagate uncertainty? With Monte Carlo framework, we make fewer assumptions about data distributions and MC allows us to preserve full data distributions (or pdfs) instead of approximating distributions. Additionally, it is straightforward to link uncertainty from different sources. The primary downsides to the MC approach include the intensity of the data and computational requirements. For example, to process 300 million ha of forest with 1 ha pixels at 2 bytes per pixel would be 600 MB / layer. If we conduct 1,000 iterations, 600 GB are required per product. As our project progresses, several issues need to be continually addressed. We proposed a "complete" accounting of uncertainty. Working with the user community and our co-investigators, we are in the process of identifying and agreeing on the definitional boundaries (i.e. *which sources of uncertainty to include*). A close collaborative approach is required and models and data need to be freely shared. This approach is most accurate when models are run multiple times, requiring input from collaborators.