International Journal of Applied Earth Observation and Geoinformation 31 (2014) 13–24

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag

# Fusion of pan-tropical biomass maps using weighted averaging and regional calibration data

Yong Ge[a,*], Valerio Avitabile[b], Gerard B.M. Heuvelink[c], Jianghao Wang[a], Martin Herold[b]

[a] State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences & Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[b] Laboratory of Geo-information Science and Remote Sensing, Wageningen University, Wageningen, The Netherlands
[c] Soil Geography and Landscape group, Wageningen University, Wageningen, The Netherlands

ARTICLE INFO

*Article history:*
Received 14 March 2013
Accepted 21 February 2014
Available online 27 March 2014

*Keywords:*
Biomass maps
Accuracy
Fusion model
Spatial variability

ABSTRACT

Biomass is a key environmental variable that influences many biosphere–atmosphere interactions. Recently, a number of biomass maps at national, regional and global scales have been produced using different approaches with a variety of input data, such as from field observations, remotely sensed imagery and other spatial datasets. However, the accuracy of these maps varies regionally and is largely unknown. This research proposes a fusion method to increase the accuracy of regional biomass estimates by using higher-quality calibration data. In this fusion method, the biases in the source maps were first adjusted to correct for over- and underestimation by comparison with the calibration data. Next, the biomass maps were combined linearly using weights derived from the variance–covariance matrix associated with the accuracies of the source maps. Because each map may have different biases and accuracies for different land use types, the biases and fusion weights were computed for each of the main land cover types separately. The conceptual arguments are substantiated by a case study conducted in East Africa. Evaluation analysis shows that fusing multiple source biomass maps may produce a more accurate map than when only one biomass map or unweighted averaging is used.



## 1. Introduction

Biomass data are important for assessing efforts in reducing carbon emissions from deforestation and forest degradation and provide valuable information in support of the conservation and sustainable management of forests, which are often missing in developing countries in particular (Baccini et al., 2012; De Sy et al., 2012; Herold and Skutsch, 2011; Romijn et al., 2012). Accurate estimation of biomass is also essential for monitoring the global or regional carbon cycle (Houghton, 2005; Le Quere et al., 2009). Currently, three methods are used for estimating the spatial distribution of biomass (Avitabile et al., 2011; Goetz et al., 2009; Sales et al., 2007). The first method uses estimates based on the assessment of different land uses and their associated biomass densities (Avitabile et al., 2011). The second method interpolates or extrapolates biomass estimates obtained in the field across extensive regions (Sales et al., 2007). The third method makes use of empirical, parametric and non-parametric models for biomass estimation over large areas using remote sensing data and spatial datasets such as land cover data, digital elevation models (DEMs) and forest maps (Avitabile et al., 2012; Baccini et al., 2012; Saatchi et al., 2011). Biomass maps can be produced at regional and global scales with different mapping methods and using a variety of input data such as field (plot) data, remotely sensed images and spatial datasets. Different biomass estimation methods may lead to different biomass maps with different accuracies that can vary from region to region (Avitabile et al., 2011; De Sy et al., 2012). The information content of multiple biomass maps may also be disparate and complementary. Fusing these maps might lead to a more accurate combined biomass map for regional and global applications. However, most current studies are concerned with producing biomass maps using a variety of methods and comparing biomass maps derived from each information source separately (Avitabile et al., 2011; Sales et al., 2007; Wulder et al., 2008). Little to no research seems to address the fusion of biomass maps from different sources using additional reference data and understanding of uncertainties in the various maps (De Sy et al., 2012; Romijn et al., 2012).

Biomass map fusion could be accomplished with existing data fusion theories and methods, which mainly include Bayesian decision theory (Mascarenhas et al., 1996), Shafer's theory of evidence

* Corresponding author. Tel.: +86 10 64888053; fax: +86 10 64889630.
*E-mail addresses:* gey@lreis.ac.cn, gey@igsnrr.ac.cn (Y. Ge).

http://dx.doi.org/10.1016/j.jag.2014.02.011

(Lee et al., 1987), consensus theory (Benediktsson and Swain, 1992), neural networks (Benediktsson et al., 1990; Bruzzone, 1999), Markov random field (Solberg et al., 1996), support vector machines (Waske and Benediktsson, 2007), fuzzy logic (Stover et al., 1996) and other statistical approaches (Bates and Granger, 1969; Hansen, 2008; Heuvelink and Bierkens, 1992; Ramin et al., 2012; Stein, 2005). Some of these methods are restricted to categorical source data while others apply to continuous source data. For example, Bayesian decision theory, Shafer's theory of evidence and consensus theory were used to fuse the classified results from different classifiers (Ge and Bai, 2010). Fuzzy logic was used to fuse land cover products by exploring disagreements among them and capturing the uncertainties in the products (Fritz and See, 2005; Herold et al., 2006; Jung et al., 2006; See and Fritz, 2006), while some statistical techniques such as principal component analysis and regression can be used to fuse continuous source data like remote sensing imagery (Hall, 1992; Pohl and van Genderen, 1998). Recently, Nguyen et al. (2012) proposed a geostatistical-based data fusion for predicting the aerosol process from two noisy datasets. This method is applicable to the process which has a linear trend in the spatial covariates, and it involves the inversion of large-to-massive datasets with fixed rank kriging.

For continuous biomass map fusion, this article adopts an idea initially proposed by Bates and Granger (1969) and Heuvelink and Bierkens (1992), which achieves biomass map fusion by weighted averaging of the source biomass maps. The weights are derived from the variance–covariance matrix associated with the errors in each of the source maps. Compared with other fusion methods, the proposed method is simple and easy to implement and complements regional estimates where additional data are available. The variance–covariance matrix is estimated by comparison of the source map values with those of a more accurate map, which must be available for a subset of the entire study area. To investigate the performance of this method, three biomass maps were fused for the study area of East Africa. The fused result was first compared with field biomass observations and the uncertainty map of fused result was then compared with the source uncertainty maps.

## 2. Methods

### 2.1. The generic framework

This article considers the case in which there are multiple maps of the same environmental variable that each has their own degree of accuracy. The aim is to fuse or merge these maps ('source maps') in such a way that the combined map has greater accuracy than each of the individual maps. The fusion method proposed here has six steps (Fig. 1): (1) collecting data, such as the source maps, calibration data and validation data; (2) pre-processing, including matching the projections and spatial resolutions of each of the source maps; (3) stratification of the study area into subareas that are homogeneous with respect to the accuracy of the source maps; (4) assessing the accuracy of the source maps per sub-area with a bias estimate and a variance–covariance matrix; (5) the fusion itself; (6) evaluation. These steps are described in detail in the sections below.

### 2.2. The fusion model

Let there be $p$ source maps $z_i$, $i = 1, \ldots, p$ that each contains estimates of the environmental variable of interest (i.e. biomass) for locations $s \in \mathrm{D}$, D being the geographical domain of interest. The fusion of these maps is achieved by a weighted linear average:

$$f(s) = \sum\nolimits_{i=1}^{p} w_i(s) \cdot (z_i(s) - v_i(s)) \tag{1}$$

where $f$ is the fused map, the $w_i(s)$ is the weight and the $v_i(s)$ is the bias correction. In order to calculate the weights we assume a statistical model for the discrepancies between the true biomass $b$ and the map estimates $z_i$:

$$z_i(s) = b(s) + v_i(s) + \varepsilon_i(s) \tag{2}$$

where $\varepsilon_i(s)$ is a random noise term with zero mean for each $s \in \mathrm{D}$. We further assume that the $\varepsilon_i(s)$, $i = 1, \ldots, p$ are jointly normally distributed with variance-covariance matrix $\mathbf{C}(s)$.

Under these assumptions, the vector of weights $\mathbf{w}(s)$ that minimizes the variance of the estimation error of $f(s)$ is easily calculated as (Searle, 1971; Heuvelink and Bierkens, 1992):

$$w(s)^{\mathrm{T}} = (1^{\mathrm{T}}\mathbf{C}(s)^{-1}1)^{-1} 1^{\mathrm{T}}\mathbf{C}(s)^{-1} \tag{3}$$

where $1 = [1, \ldots, 1]^{\mathrm{T}}$ is the $p$-dimensional unit vector and where T means transpose. The fused map will be unbiased with variance given by:

$$\mathrm{Var}(f(s)) = (1^{\mathrm{T}}\mathbf{C}(s)^{-1}1)^{-1} \tag{4}$$

Under the assumptions made, the variance of the error in the fused map cannot be greater than the smallest of the error variances of the individual maps (Bates and Granger, 1969). It will be substantially smaller than that if multiple maps have similar error variances that are close to the smallest error variance and the errors associated with these maps are not strongly positively correlated. Attractive properties of this fusion model are: (1) it is simple to use; (2) it allows the assignment of different weights to different biomass maps based on the accuracies of the maps for different strata (i.e. subareas); and (3) it can handle missing data values in some of the source maps. The latter means that when one or more (but not all) source maps have missing data at some location, then the fused result at that location is obtained by fusing the remaining source maps, with weights derived from the reduced variance–covariance matrix.

### 2.3. Calibration of the fusion model

From the above equations it can be seen that to compute the weights, the variances of the source map errors are needed, as well as their correlations (or covariances). In addition, bias estimates for each of the individual source maps are required. In biomass fusion, one way to obtain this information is to use calibration data. However, accessing calibration data which cover the entire area of interest is difficult. An alternative method for deriving the accuracies of source maps and their correlations is to use calibration data from a subarea for which a biomass map with much higher accuracy is available. By comparison of the calibration and map data, one obtains the errors of the maps, and from these variances, correlations and biases can be estimated in the usual way.

Since it is unrealistic to assume that the variances, correlations and biases of the source maps are spatially invariant, it is sensible to stratify the area into zones for which this assumption is more realistic. Indeed, the spatial accuracy of the source biomass maps may vary from region to region. Given calibration data $y(s_k)$ at $k = 1, \ldots, n$ locations in a stratum, the estimates of bias and variance-covariance matrix are as follows:

$$\hat{v}_i = \frac{1}{n}\sum_{k=1}^{n}(z_i(s_k) - y(s_k)) \tag{5}$$

$$\hat{\mathbf{C}}_{ij} = \frac{1}{n}\sum_{k=1}^{n}(z_i(s_k) - y(s_k))(z_j(s_k) - y(s_k)) \tag{6}$$

where $i, j = 1, \ldots, p$ indicate source maps.