Remote Sensing of Environment 203 (2017) 240–255

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Stratification and sample allocation for reference burned area data

Marc Padilla [a,*], Pontus Olofsson [b], Stephen V. Stehman [c], Kevin Tansey [a], Emilio Chuvieco [d]

[a] Centre for Landscape and Climate Research, Department of Geography, University of Leicester, UK
[b] Department of Earth and Environment, Boston University, USA
[c] Department of Forest and Natural Resources Management, College of Environmental Science and Forestry, State University of New York, Syracuse, NY, 13210, USA
[d] Environmental Remote Sensing Research Group, Department of Geology, Geography and Environment, Universidad de Alcalá, Spain

ARTICLE INFO

Article history:
Received 30 September 2016
Received in revised form 26 June 2017
Accepted 29 June 2017
Available online 8 July 2017

Keywords:
Validation
Probability sampling
Terrestrial globe
Fire Disturbance
Sampling efficiency

ABSTRACT

Statistical estimation protocols are one of the key means to ensure that independent and objective information on product accuracy is communicated to end-users. Methods for validating burned area products have been developed based on a probability sample of a space by time partitioning of the population. We extend this basic methodology to improve stratification and sample allocation, key elements of a sampling design used to collect burned area reference data. We developed and evaluated an approach to partition each year and biome into low and high burned area (*BA*) strata. Because the threshold used to separate the sampling units into low and high *BA* can vary by year and biome, this approach offers a more targeted stratification than used in previous studies for which a common threshold was applied to all biomes. A hypothetical population of validation data was then used to quantitatively compare the precision of accuracy estimates derived from different stratification and sample size allocation options. We evaluated two options that had been previously examined in the *BA* validation literature, and extended previous studies by adding two new options specifically developed for ratio estimates. Stratification based on mapped *BA* reduced standard errors of the global burned area accuracy estimates from one-half to one-eighth relative to standard errors of simple random sampling. Stratifying by mapped *BA* was also found to reduce standard errors of accuracy estimates for most year by biome strata indicating that this advantage of stratification and sample allocation applies generally to a range of conditions (i.e., biomes and years). The most precise estimates were obtained using a sample size per stratum allocation $n_h \propto N_h \sqrt{\overline{BA}_h}$ where N_h is the number of units in stratum h and $\overline{BA}_h$ is the mean mapped *BA* for stratum h. The best sampling design from our analyses was then used to select a set of 1,000 samples from a hypothetical population of validation data and confidence intervals were computed for each sample. Close to 95% of these confidence intervals contained the true population value thus confirming the validity of confidence intervals produced from the estimates and standard errors.

1. Introduction

Biomass burning is one of the most important processes impacting the Earth system (Bond and Keeley, 2005; Bowman et al., 2009) and one of the main sources of gases and aerosols emitted to the atmosphere (van der Werf et al., 2004, 2010). The Global Climate Observing System (GCOS) program identified Fire Disturbance as an Essential Climate Variable (ECV) (GCOS, 2004), commonly expressed by burned area (*BA*) information (Mouillot et al., 2014). Global *BA* products provide the location and dates of burned surfaces at a coarse spatial resolution (300–1000 m).

Product validation is defined as "… the process of assessing, by independent means, the quality of the data products derived from the system outputs" (CEOS-WGCV, 2012). *BA* products usually cover multi-year periods and the Committee on Earth Observation Satellites (CEOS) Land Product Validation Subgroup (LPV) highlights the importance of assessing the temporal stability of a product's accuracy by collecting data over globally representative locations and time periods (http://lpvs.gsfc.nasa.gov). The selection of representative samples is particularly important when the event to be characterized is rare and occurs in spatio-temporal clusters, e.g. fires (Chou et al., 1993; Giglio et al., 2010). The main challenge is to define an optimal sampling design that leads to precise accuracy estimates and allocates the sample through several time periods and regions of interest, e.g. years in a multiyear time period and major biomes. Throughout the manuscript optimal sampling design refers to the design that minimizes the variance of an accuracy estimate, for a specific sample size (Cochran, 1977; Section 5.5). In our application, the factors evaluated that can affect the optimal design are the strata and the allocation of the sample to these strata.

[*] Corresponding author.
E-mail address: mp489@le.ac.uk (M. Padilla).

http://dx.doi.org/10.1016/j.rse.2017.06.041
0034-4257/

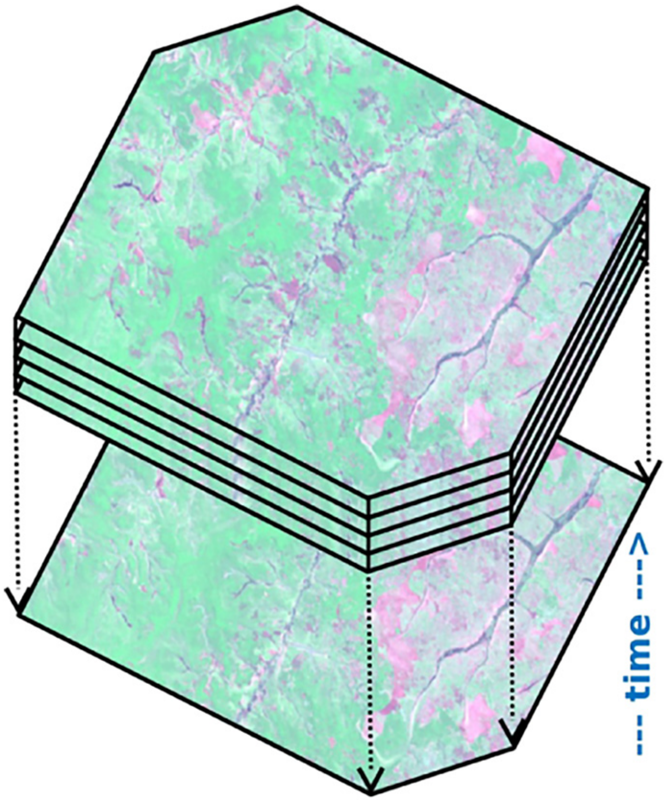

Fig. 1. Illustration of the "voxel" sampling units proposed by Boschetti et al. (2016) for partitioning the three-dimensional space by time population. Each sampling unit is delimited spatially by a Thiessen Scene Area (TSA) partitioning the two dimensions of space and temporally (the third dimension) by the time between two consecutive Landsat images. The image at the bottom is displaced further down to illustrate temporal spectral changes. Images are displayed as false color composites with SWIR, NIR and red bands in the red, green and blue channels respectively.

BA reference data are commonly obtained following the recommendations of the CEOS LPV (Boschetti et al., 2009), using multi-temporal pairs of medium spatial resolution images (10–60 m). Methodological limitations and reference data generation costs led until recently to relatively small sample sizes, the selection of sites being based on expert knowledge (Chuvieco et al., 2008; Giglio et al., 2009; Padilla et al., 2014b; Plummer et al., 2007; Roy and Boschetti, 2009; Tansey et al., 2008). Recently, global accuracy estimates were first produced from a probability sampling design consisting of a spatially stratified random sample (Padilla et al., 2014a, 2015). Boschetti et al. (2016) developed a sampling approach based on partitioning the space by time population into three-dimensional "voxel" units defined by Thiessen Scene Areas (TSAs) to partition space and 16-day Landsat image pairs to partition time. Although the specific example presented by Boschetti et al. (2016) used TSAs and a time period specific to Landsat, the general approach using the voxel units can be applied to other spatial and temporal partitions. Boschetti et al. (2016) also proposed a stratification based on a threshold of active fire counts that split each biome into two strata representing low and high fire activity. For all geographic strata (biomes), Boschetti et al. (2016) determined this threshold using the 20th percentile of the cumulative distribution of active fire counts. Boschetti et al. (2016) acknowledged that additional work was needed to investigate "the impact of different thresholds to define low and high fire activity." Furthermore, the allocation methods they analysed did not include two methods recommended by Cochran (1977; Section 6.14) for ratio estimates. The main accuracy measures for burned area are in fact ratios, as is for example the case of the commission and omission error rates.

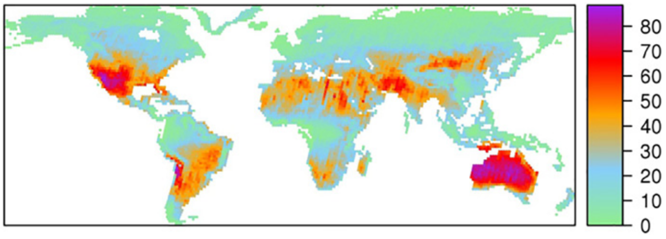

Fig. 2. Spatial distribution of reference data availability for 2003 to 2014 expressed as percentage of time that Thiessen Scene Areas covered by Landsat image pairs separated by 16 days or fewer are available from the USGS archive (accessed on September 2015).

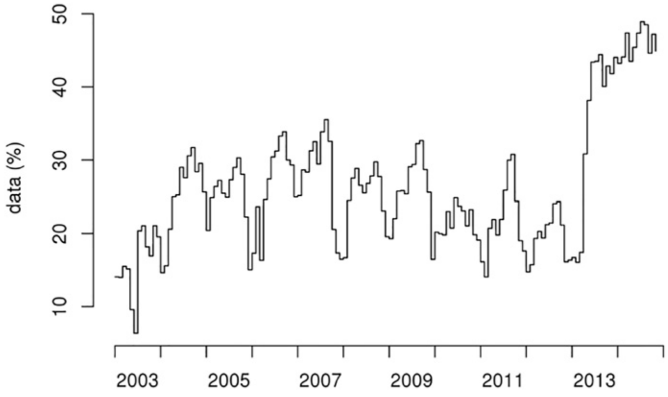

Fig. 3. Temporal distribution of reference data availability expressed as monthly percentage of area·time covered by Landsat image pairs separated by 16 days or fewer.

The objective of the current study is to improve the sampling design by: (a) allowing the threshold used to define low and high burned area strata to vary by biome and year, and (b) identifying the best method for allocating the sample to strata among those methods recommended in the literature. The precision of the accuracy estimates obtained for the different stratification and sample allocation methods was compared using a hypothetical population of validation data. The hypothetical population allows for direct comparison of the standard errors of accuracy estimates for the different design options evaluated. These comparisons provide quantitative information to guide decisions regarding how to construct strata separating low and high burned area and how to effectively allocate the sample to these strata.

The current research is in the framework of the validation effort of the Fire Disturbance (Fire_cci) Phase II project (www.esa-fire-cci.org), part of the European Space Agency's (ESA) Climate Change Initiative. The main goal of this effort is to generate reference data that cover the twelve year period 2003–2014. This period was defined by the time period for which data were available for both global sensors used in the project, MEdium Resolution Imaging Spectrometer (MERIS) and Moderate Resolution Imaging Spectroradiometer (MODIS).

Table 1
Population error matrix for a single Thiessen Scene Area (TSA) sampling unit. Matrix cells express agreement (diagonal cells) or disagreement (off-diagonal cells) in terms of area (m^2) between the *BA* product (map) class and the reference classification of the entire unit.

Product classification	Reference classification		Row total
	Burned	Unburned	
Burned	E_{11}	E_{12}	E_{1+}
Unburned	E_{21}	E_{22}	E_{2+}
Col. Total	E_{+1}	E_{+2}	

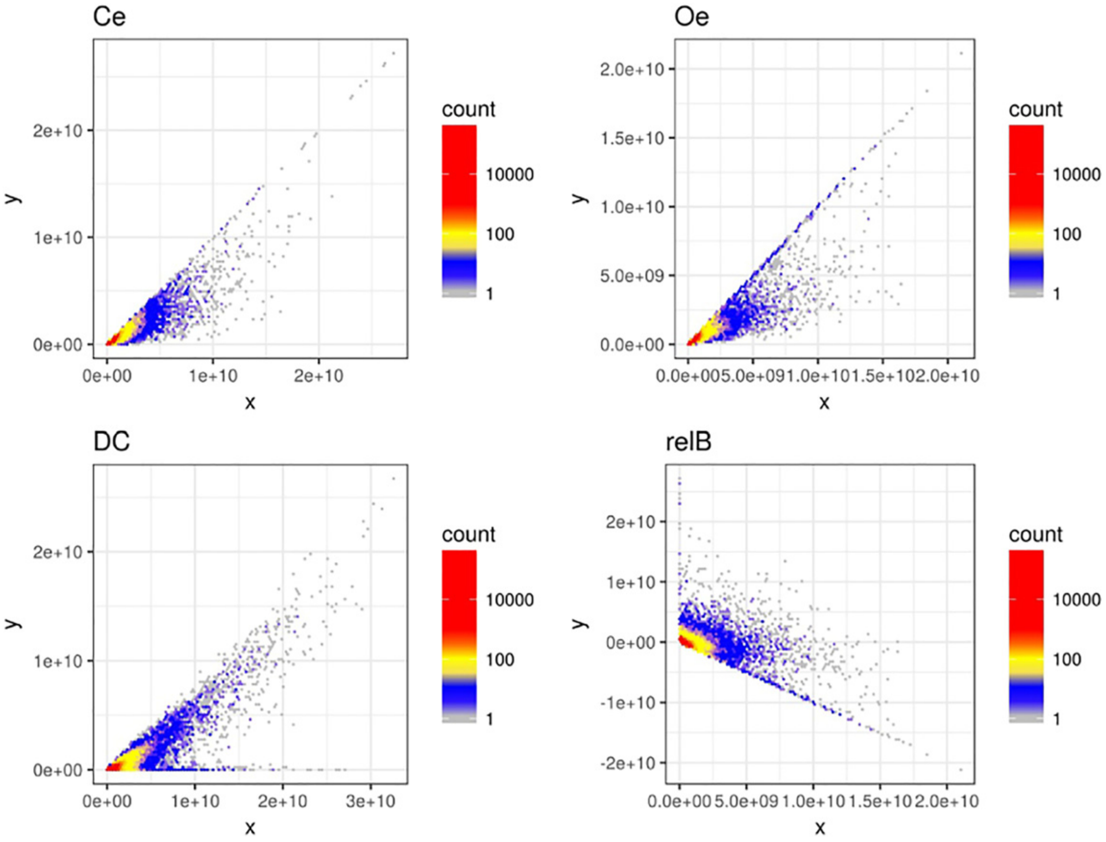

Fig. 4. Scatter plots between y_i and x_i for *Ce*, *Oe*, *DC* and *relB* for the hypothetical population of validation data.

Sections 2 and 3 of the article present the methodology and results respectively. Section 2.1 presents the definition of the voxel sampling units, Section 2.2 documents the equations of the accuracy estimates, and Section 2.3 describes the options for the variable used to determine stratum boundaries and the sample allocation to strata. Section 2.4 presents the method for choosing a threshold to separate low and high *BA* strata, and Section 2.5 describes the hypothetical population used to quantify the precision of the different stratification and sample allocation options defined in Sections 2.3 and 2.4. Section 2.6 completes the methodology by describing a study to evaluate the validity of the confidence intervals for accuracy measures estimated using the proposed sampling design and sample size feasible for the Fire_cci project. Section 3.1 of the Results presents the findings regarding the threshold for separating low and high *BA* stratum and the sample allocation to strata for the global sampling design. Section 3.2 shows the results of the precision comparison for the different design options for the global accuracy estimates and Section 3.3 presents the same precision comparisons at the individual biome level by year. The Results conclude with the confidence interval coverage properties presented in Section 3.4. Sections 4 and 5 present the discussion and conclusions of the implications of the findings.

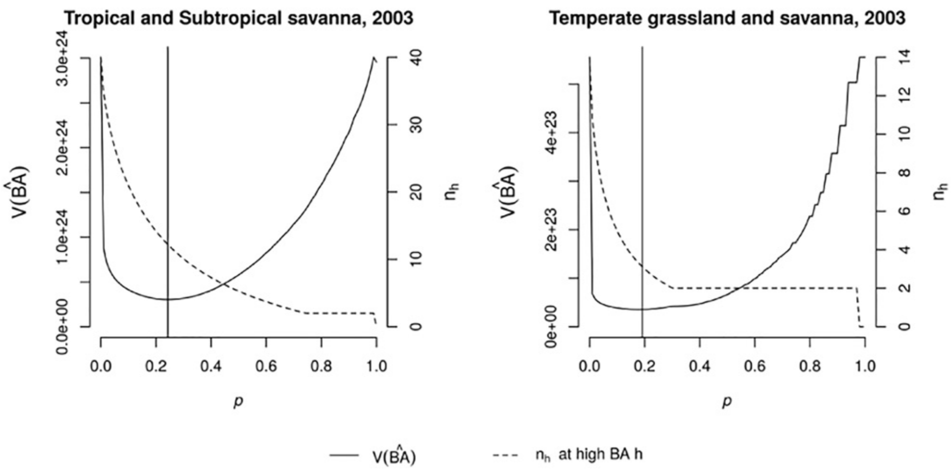

Fig. 5. Examples of selection of an optimal threshold p^* that divides a year-biome stratum into two parts to minimize $V(\widehat{BA})$ while ensuring $n_h \geq 2$. Candidate thresholds cover the range of the cumulative sum of the BA_i distribution relative to the total year-biome *BA*, from 0 to 1 with increments of 0.01. Vertical lines indicate the optimal thresholds p^* selected.

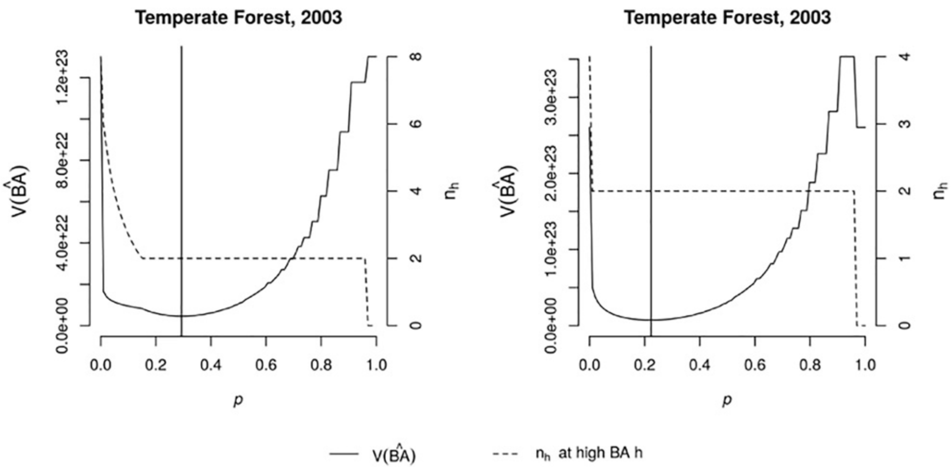

Fig. 6. As in Fig. 5 but for 2003 Temperate Forests and allocation methods $n_h \propto N_h \sqrt{\overline{BA}_h}$ (left) and $n_h \propto N_h \overline{BA}_h$ (right).

2. Methods

2.1. Sampling units

As in Padilla et al. (2014a, 2015) and Boschetti et al. (2016), the spatial dimension of sampling units was based on Landsat World Reference System II (WRS-II) to simplify data downloading and processing. The spatial dimension of the sampling units was defined by the Thiessen Scene Areas (TSAs) constructed by Cohen et al. (2010) and Kennedy et al. (2010) specifically for use with Landsat WRS-II frames. The key advantage of TSAs is that they partition the spatial domain into non-overlapping Landsat-like frames, which allow for a convenient computing of unbiased estimators (Gallego, 2005).

Following the CEOS Validation protocol for *BA* products (Boschetti et al., 2009), reference data are generated from two consecutive images acquired for the same TSA. Therefore, a sampling unit is delimited spatially by a TSA and temporally by the acquisition dates of two consecutive images (an image pair) (Fig. 1). For the analyses presented in this article, an image pair forms a sampling unit whenever the pair is separated by 16 days or less (the time unit can be defined based on the imagery used). It is relevant to limit the time length between image pairs to ensure the spectral signal of a fire that occurred between acquisition times is still present in the latest image. The duration of a fire spectral signal can be very short particularly for grasslands with wet soils.

Landsat imagery with <30% of clouds at the USGS archive (http://landsat.usgs.gov/ on September 2015) and the temporal requirement between image pairs specified above limits the availability of reference data. Globally from 2003 to 2014 only 26.24% of the area·time is covered by the image pairs available at the USGS archive. Fig. 2 shows the spatial distribution of such availability which appears to be affected by global cloud coverage patterns and by Landsat archiving strategies. Fig. 3 shows the temporal distribution of reference data availability with clear periodic peaks in the middle of the year and a large increase from 2013 onwards, produced from the start of the Landsat 8 campaign.

2.2. Accuracy estimates

Commonly in *BA* validation, accuracy estimates are based on the cross tabulation approach (Congalton and Green, 1999; Latifovic and Olthof, 2004) as summarized by an error matrix which quantifies agreement and disagreement in terms of area (m^2) between product and reference classifications (Table 1). For both the product and reference classification, a pixel is coded as "burned" if fire is detected between the dates defining the temporal dimension of the sampling unit, "unburned" if fire is not detected, or "no-data" for unobserved pixels (e.g., due to cloud coverage).

Accuracy measures are commonly ratios of combinations of error matrix entries. For example, for the "burned" class the Commission error ratio is

$$Ce = \frac{E_{12}}{E_{1+}} \tag{1}$$

and the omission error ratio is

$$Oe = \frac{E_{21}}{E_{+1}} \tag{2}$$

where E_{ij} refers to the population values of the error matrix entries (Table 1). Recent publications (Padilla et al., 2014a,b, 2015) also used the Dice coefficient (*DC*) (Dice, 1945) and measures of bias. *DC* is particularly useful when comparing product accuracies as it combines both error ratios (*Ce* and *Oe*) into a single summary measure of accuracy of the category "burned". *DC* has a sensible probabilistic interpretation

Table 2
Optimal *BA* thresholds p^* for allocation $n_h \propto N_h \sqrt{\overline{BA}_h}$.

	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Tropical and Subtropical savanna	0.24	0.26	0.29	0.26	0.26	0.28	0.27	0.28	0.29	0.30	0.26	0.24
Tropical Forest	0.39	0.28	0.43	0.36	0.29	0.35	0.41	0.28	0.36	0.37	0.27	0.39
Temperate grassland and savanna	0.19	0.16	0.25	0.21	0.23	0.23	0.28	0.25	0.27	0.30	0.17	0.22
Boreal Forest	0.25	0.14	0.15	0.16	0.14	0.19	0.14	0.16	0.16	0.20	0.18	0.18
Temperate Forest	0.29	0.18	0.25	0.36	0.24	0.29	0.29	0.29	0.28	0.30	0.26	0.24
Mediterranean Forest	0.21	0.23	0.22	0.22	0.22	0.18	0.18	0.21	0.24	0.17	0.17	0.19
Others	0.15	0.18	0.31	0.13	0.16	0.38	0.30	0.28	0.12	0.15	0.13	0.14

Table 3
Optimal *BA* thresholds p^* for allocation $n_h \propto N_h \overline{BA}_h$.

	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Tropical and Subtropical savanna	0.14	0.14	0.17	0.15	0.15	0.18	0.14	0.16	0.17	0.17	0.14	0.13
Tropical Forest	0.19	0.18	0.20	0.19	0.18	0.21	0.22	0.17	0.21	0.21	0.18	0.20
Temperate grassland and savanna	0.13	0.15	0.15	0.14	0.16	0.13	0.16	0.19	0.17	0.16	0.13	0.14
Boreal Forest	0.17	0.14	0.15	0.16	0.14	0.16	0.14	0.16	0.16	0.17	0.13	0.12
Temperate Forest	0.22	0.18	0.22	0.31	0.21	0.23	0.23	0.25	0.25	0.28	0.22	0.21
Mediterranean Forest	0.21	0.23	0.22	0.22	0.22	0.18	0.18	0.21	0.24	0.17	0.17	0.19
Others	0.14	0.11	0.17	0.08	0.13	0.22	0.17	0.13	0.06	0.08	0.14	0.09

(Dice, 1945; Fleiss, 1981; Forbes, 1995; Hand, 1981; Hellden, 1980; Liu et al., 2007) as it is the conditional probability that one classifier identifies a pixel as burned, given that the other classifier also identified it as burned (Fleiss, 1981).

$$DC = \frac{2E_{11}}{2E_{11}+E_{12}+E_{21}} \tag{3}$$

Bias is of interest to end-users (Mouillot et al., 2014) as it quantifies the difference between the total area burned detected by the map versus the reference classification,

$$bias = E_{1+} - E_{+1} = E_{12} - E_{21} \tag{4}$$

It is also useful to express bias relative to the reference *BA*, so relative bias is defined as

$$relB = \frac{E_{12}-E_{21}}{E_{+1}} \tag{5}$$

Global estimates of accuracy are computed taking into account the stratified sampling design and using a stratified combined ratio estimator (Cochran, 1977; Section 6.11) of the form

$$\hat{R} = \frac{\hat{Y}}{\hat{X}} = \frac{\sum_{h=1}^{H} N_h \overline{Y}_h}{\sum_{h=1}^{H} N_h \overline{X}_h} \tag{6}$$

where H is the number of strata, N_h is the number of sampling units in stratum h, $\overline{Y}_h$ and $\overline{X}_h$ are the means of y_i and x_i at stratum h, and y_i and x_i are values defined for sample unit i based on the denominator and numerator of the different accuracy measures: for *Ce*, $y_i = E_{12}$ and $x_i = E_{1+}$; for *Oe*, $y_i = E_{21}$ and $x_i = E_{+1}$; for *DC*, $y_i = 2E_{11}$ and $x_i = 2E_{11}+E_{12}+E_{21}$; and for *relB*, $y_i = E_{12} - E_{21}$ and $x_i = E_{+1}$. To simplify notation, we do not include a stratum subscript "h" and we suppress the subscript "i" for the Table 1 error matrix values that are associated with sample unit *i*.

The variance of the ratio estimator $\hat{R}$ is

$$V\left(\hat{R}\right) = \frac{1}{X^2}\sum_{h=1}^{H} N_h^2\left(1-\frac{n_h}{N_h}\right)S_{uh}^2/n_h \tag{7}$$

$$S_{uh}^2 = \frac{1}{N_h-1}\sum_{i=1}^{N_h}\left(u_i-\overline{U}_h\right)^2 \tag{8}$$

where $\overline{U}_h$ is the mean of u_i for stratum h, $u_i = y_i - Rx_i$, and X is the population total (over all strata) of x_i.

The bias (Eq. (4)) and the *BA* based on the reference data (denoted *BAref*), are expressed as population total estimates of the form

$$\hat{Y} = \sum_{h=1}^{H} N_h \overline{y}_h \tag{9}$$

where $\overline{y}_h$ is the sample mean of y_i in stratum h, and y_i is defined by $E_{12} - E_{21}$ for estimating bias and y_i is defined as E_{+1} for estimating *BAref.*

The variance of the estimated total $\hat{Y}$ is

$$V\left(\hat{Y}\right) = \sum_{h=1}^{H} N_h^2\left(1-\frac{n_h}{N_h}\right)S_{yh}^2/n_h \tag{10}$$

$$S_{yh}^2 = \frac{1}{N_h-1}\sum_{i=1}^{N_h}\left(y_i-\overline{Y}_h\right)^2 \tag{11}$$

As noted in the previous Section, reference data are not always available. For the analyses presented in this article, the hypothetical population used to compare variance of different stratification and sample allocation options was defined on the basis of available reference data to construct the population.

2.3. Options for auxiliary variable used to define strata and optimum sample allocation

For a probability sampling design that employs stratification, the sample size n_h from each stratum may be chosen to minimize the variance of estimates. For example, the variance of a population total estimate ($\hat{Y}$; e.g. *BAref*) is minimized if the sample size n_h is proportional to $N_h S_{yh}$ (Cochran, 1977; Section 5).

$$n_h \propto N_h S_{yh} \tag{12}$$

The variance of a ratio estimator ($\hat{R} = \hat{Y}/\hat{X}$; e.g. all accuracy measures) depends on the deviations $u_i = (y_i - Rx_i)$ for each stratum. Specifically, the variance of $\hat{R}$ is minimized if the sample size n_h is

Table 4
Table with the sample sizes n_h for each year (columns), biome (rows) and *BA* level (high *BA* on the left of the "+" sign and low *BA* on the right) for allocation $n_h \propto N_h \sqrt{\overline{BA}_h}$.

	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Tropical and Subtropical savanna	13 + 30	15 + 33	13 + 33	14 + 30	15 + 31	13 + 31	12 + 32	11 + 33	11 + 30	13 + 30	17 + 34	16 + 30
Tropical Forest	2 + 13	3 + 12	2 + 14	2 + 11	3 + 11	2 + 10	2 + 14	3 + 12	2 + 10	2 + 10	3 + 10	2 + 12
Temperate grassland and savanna	3 + 9	2 + 7	2 + 9	3 + 9	2 + 9	3 + 10	2 + 10	2 + 9	2 + 10	2 + 9	2 + 7	2 + 9
Boreal Forest	2 + 6	2 + 2	2 + 2	2 + 3	2 + 2	2 + 3	2 + 2	2 + 3	2 + 3	2 + 3	2 + 5	2 + 4
Temperate Forest	2 + 6	2 + 3	2 + 3	2 + 5	2 + 4	2 + 6	2 + 5	2 + 4	2 + 4	2 + 4	2 + 4	2 + 4
Mediterranean Forest	2 + 2	2 + 2	2 + 2	2 + 2	2 + 2	2 + 2	2 + 2	2 + 2	2 + 2	2 + 2	2 + 2	2 + 2
Others	2 + 8	3 + 12	2 + 12	4 + 11	3 + 12	2 + 12	2 + 11	2 + 13	5 + 15	5 + 14	2 + 8	2 + 11

Table 5
As in Table 4 but for allocation $n_h \propto N_h \overline{BA}_h$.

	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Tropical and Subtropical savanna	59 + 10	58 + 10	57 + 12	56 + 10	56 + 10	56 + 12	59 + 10	57 + 11	50 + 11	52 + 11	60 + 10	58 + 8
Tropical Forest	5 + 2	5 + 2	5 + 2	4 + 2	6 + 2	4 + 2	5 + 2	6 + 2	4 + 2	4 + 2	5 + 2	5 + 2
Temperate grassland and savanna	6 + 2	4 + 2	5 + 2	5 + 2	5 + 2	7 + 2	5 + 2	4 + 2	4 + 2	5 + 2	4 + 2	6 + 2
Boreal Forest	2 + 2	2 + 2	2 + 2	2 + 2	2 + 2	2 + 2	2 + 2	2 + 2	2 + 2	2 + 2	2 + 2	2 + 2
Temperate Forest	2 + 2	2 + 2	2 + 2	2 + 2	2 + 2	2 + 2	2 + 2	2 + 2	2 + 2	2 + 2	2 + 2	2 + 2
Mediterranean Forest	2 + 2	2 + 2	2 + 2	2 + 2	2 + 2	2 + 2	2 + 2	2 + 2	2 + 2	2 + 2	2 + 2	2 + 2
Others	2 + 2	5 + 2	3 + 2	7 + 2	5 + 2	3 + 2	3 + 2	4 + 2	13 + 2	10 + 2	3 + 2	5 + 2

proportional to the product of the stratum size, N_h, and the standard deviation of u_i for each stratum, S_{uh} (Cochran, 1977; Section 6),

$$n_h \propto N_h S_{uh} \tag{13}$$

It is difficult to specify S_{yh} and S_{uh} at the planning stage of the sampling design because reference data have not yet had been collected and hence y_i and in some cases x_i are not available.

For ratio estimates and when x_i is available for the entire population, Cochran (1977; Section 6.14) recommends two sample allocation options depending on whether S_{uh} is expected to be proportional to $\sqrt{\overline{X}_h}$ or to $\overline{X}_h$. At the sample design planning stage, mapped *BA* is the only practical information available regarding variability so Cochran's (1977) two suggested sample allocation methods are implemented as follows using mapped *BA* as the x_i value for each element *i*:

$$n_h \propto N_h \sqrt{\overline{BA}_h} \tag{14}$$

$$n_h \propto N_h \overline{BA}_h \tag{15}$$

The *BA* product MCD64 (Giglio et al., 2009) was identified as the most accurate product in the results of Padilla et al. (2015) and its burned area extent estimates at sampling units (BA_i) were used to implement the allocation methods. Therefore, the efficiency of the allocation methods will be greater for a particular accuracy estimate as BA_i is closer to the denominator (x_i) of that accuracy measure. For example, the denominator for estimating *DC* is the sum of *BA* in the reference data and the *BA* in the product, so the effectiveness of the sample allocation will depend on how closely mapped *BA* is correlated with the denominator of *DC*. A shortcoming of using a global *BA* product is that it may miss small fires (Hantson et al., 2013; Randerson et al., 2012). If small fires are omitted and they contribute a large area, the allocation method would be less effective in terms of reducing standard errors of the estimates. This same shortcoming is described by Hansen et al. (1946) for surveys of business sales, who highlighted that such errors would not introduce bias into the sample estimates, but would diminish the variance reduction achieved by the chosen sample allocation.

2.4. Stratification and sample allocation

The stratification is constructed in three levels to allow control of the sample size by calendar year in each of the major Olson biomes (Olson et al., 2001) and by low and high fire activity:

- The first stratification level consisted of assigning each sampling unit to a calendar year. For consistency and simplicity, this assignment was based on the earliest acquisition date of the Landsat image pair. A yearly stratification is convenient as it offers flexibility when planning the data collection. In particular, this first level of stratification makes easy to extend the temporal period of study by adding complete years.
- The second stratification level consisted of assigning each sampling unit to the major biome for which the TSA had the maximum area (Fig. 7), as in Padilla et al. (2014a, 2015) and Boschetti et al. (2016).
- The third stratification level was based on thresholds of *BA*. Hansen et al. (1946) recommends for ratio estimates to stratify the population by using thresholds of x_i or another related variable. As mentioned in the previous section, x_i = mapped *BA* obtained from MCD64 for sample unit *i* was used for this purpose. For each allocation method, the optimal stratification was defined by dividing sampling units into high and low *BA* by using a threshold of *BA* specifically adapted to each year-biome stratum. Given the available sample size for each year and biome, the threshold was selected to minimize $V(\widehat{BA})$, the variance of the estimate of *BA*. Recall that *BA* from the map product is the primary data available

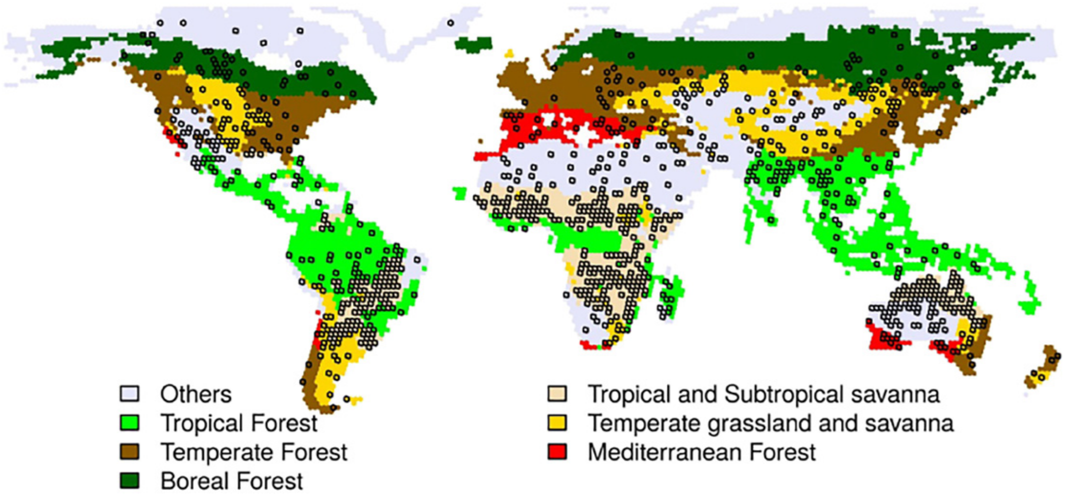

Fig. 7. Distribution of sampled Thiessen Scene Areas (TSAs) for one realization of the sample design proposed for allocation $n_h \propto N_h \sqrt{\overline{BA}_h}$. The biome stratification used on this study is based on a reclassification of the 14 Olson biomes (Olson et al., 2001).

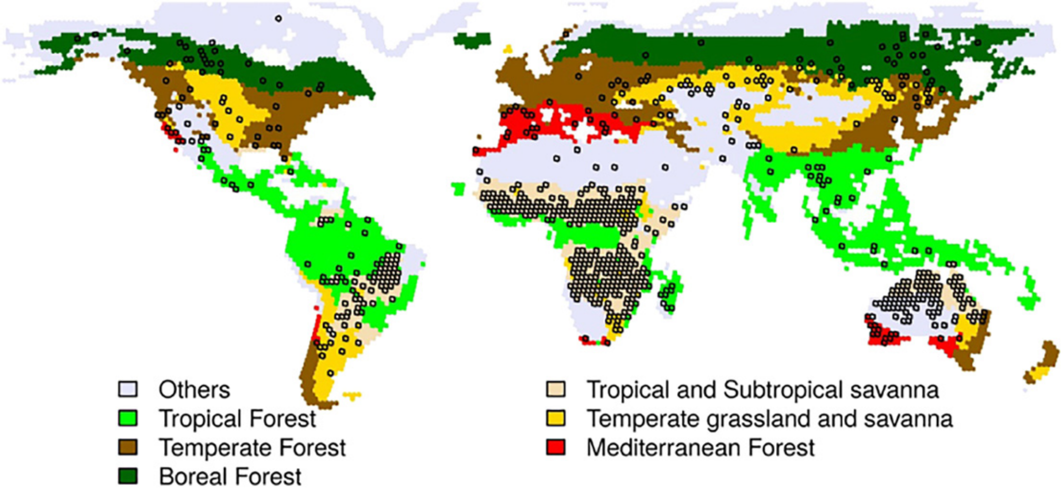

Fig. 8. Distribution of sampled Thiessen Scene Areas for one realization of the sample design proposed for allocation $n_h \propto N_h \overline{BA}_h$.

at the sample design planning stage that can be used to tailor the strata construction and sample allocation to specific estimation objectives.

The budget of the Fire_cci project allowed for 100 sampling units per year. Hence, the optimal sample size by biome for each year can be determined from the allocation methods mentioned above (Eqs. (14) and (15)) and limiting the total sample size to 100 units per year. At least two sampling units per stratum are needed to estimate the variance of the ratio estimator $\hat{R}$. Because the optimal allocation formulas do not guarantee this minimum sample size per stratum, an iterative process was used (Appendix A) to ensure that all $n \geq 4$ for all year by biome combinations while preserving as much as possible the sample allocation determined from each method. The reason for requiring $n \geq 4$ was because each biome by year stratum will ultimately be split into low and high *BA* strata so this will allow 2 sample units per stratum after the entire stratification process is complete.

To divide each biome into low and high *BA* strata, we use a threshold of *BA*. If the *BA* of sampling unit *i* is below this threshold the unit is assigned to the low *BA* stratum, and otherwise it is assigned to the high *BA* stratum. This threshold is selected to minimize $V(\widehat{BA})$ given the total sample size for each year by biome stratum and the requirement $n_h \geq 2$. The *BA* threshold used to divide a year-biome stratum into two *BA* strata then determines

- $N_h, \overline{BA}_h$ and S_{BAh} of each subsequent stratum *h* where S_{BAh} is the population standard deviation of *BA* for all units in stratum *h* (here stratum h refers to the low or high *BA* strata within a biome)
- n_h, determined according to the sample allocation method (Eqs. (14) or (15))
- $V(\widehat{BA})$, as it is function of N_h, n_h and S_{BAh} (Eq. (10), where $S_{yh} = S_{BAh}$).

For each allocation method and year and biome combination, $V(\widehat{BA})$ is estimated for a set of thresholds (denoted *p*) evenly distributed across

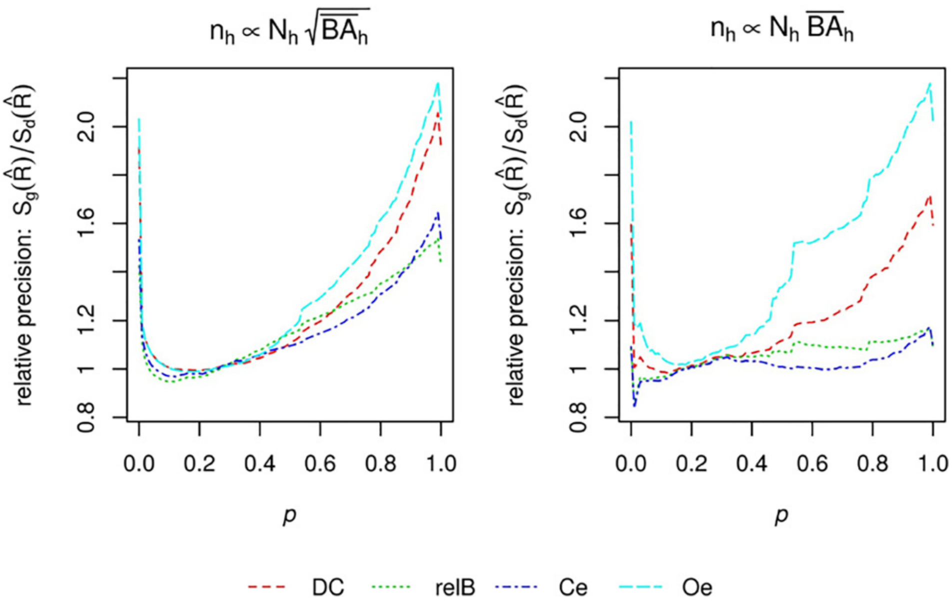

Fig. 9. Relative precision $S_g(\hat{R})/S_d(\hat{R})$ of selecting a *BA* threshold specific to each year-biome (method *d*) to selecting a global fixed *BA* threshold (method *g*) when estimating *DC*, *relB*, *Ce* and *Oe*, for allocation methods $n_h \propto N_h \sqrt{\overline{BA}_h}$ and $n_h \propto N_h \overline{BA}_h$. $S_g(\hat{R})$ was evaluated across the full range of *BA*, with *p* varied from 0 to 1 in increments of 0.01.

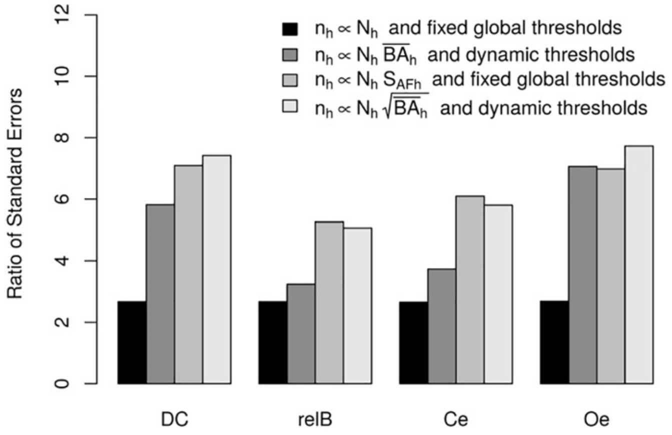

Fig. 10. Ratio of the standard errors of the accuracy estimates from simple random sampling divided by the standard errors for one of the stratified sampling options ($n_h \propto N_h \sqrt{X_h}$ with the dynamic thresholds, $n_h \propto N_h \overline{BA}_h$ with the dynamic thresholds, $n_h \propto N_h$ with global active fires (AF) $p = 0.2$ thresholds and $n_h \propto N_h S_{AFh}$ with global (AF) $p = 0.2$).

the range of the cumulative distribution percentiles from 0 to 1 in increments of 0.01. For example, $p = 0.2$ would refer to a *BA* threshold that divides a year and biome into a high *BA* stratum and a low *BA* stratum, the former containing the 20% of *BA*. That is, all elements with *BA* below the threshold established by setting $p = 0.2$ would be assigned to the low *BA* stratum. By varying *p* through the range of values between 0 and 1, the value of *p* that produces the smallest $V(\widehat{BA})$ can be determined and this optimal threshold *p* is then defined as p^*. The determination of the stratum threshold based on cumulative *BA* for the units within the biome follows the approach used by Boschetti et al. (2016).

2.5. Comparing precision of the stratification and allocation options

To compare precision of the different stratified design options, we computed the standard errors for the accuracy estimates using data from a hypothetical population. The hypothetical population included the MERIS version of the Fire_cci project (under development in the Fire Disturbance project Phase II) as the *BA* product being evaluated and the MCD45 (Roy et al., 2008) product as the reference data to validate the *BA* product. The time period covered by this population is 2005–2011. The difference between these two products represents a population (i.e., complete coverage census) that approximates a realistic pattern of classification error of *BA* products. The advantage of working with a hypothetical population with complete coverage is that it allows for evaluating many stratification and sample allocation options. Further, because we have the population we can compute standard errors for each of these options and these standard errors are "exact" in the sense of not being subject to sampling variability.

Table 7
As in Table 6 but for omission error ratio (*Oe*).

	2005	2006	2007	2008	2009	2010	2011
Tropical and Subtropical savanna	0.88	1.12	0.95	1.28	1.13	0.96	0.99
Tropical Forest	1.41	0.81	1.00	1.51	1.15	1.30	1.33
Temperate grassland and savanna	1.82	1.88	1.69	1.78	1.65	1.61	1.31
Boreal Forest	3.36	2.20	1.86	2.05	1.62	3.01	1.72
Temperate Forest	0.84	1.63	2.15	2.02	0.89	1.87	1.78
Mediterranean Forest	3.56	2.14	3.44	3.59	2.69	3.32	2.62
Others	1.77	1.96	1.58	1.63	1.96	1.41	1.57

This hypothetical population satisfies the conditions under which the ratio estimators for several of the accuracy measures are best linear unbiased estimators (Cochran, 1977; Section 6.7): (1) "The relation between y_i and x_i is a straight line through the origin" and (2) "The variance of y_i about this line is proportional to x_i". Fig. 4 shows the scatter plots between y_i and x_i for four accuracy measures, *Ce*, *Oe*, *DC* and *relB*, and how the two specified conditions are clearly met for the first three measures, but not for *relB*. Very similar distributions have been observed for samples of validation data derived from real reference data (see Appendix B for validation sample data presented by Padilla et al. (2015)). According to Cochran (1977; p. 172), the relationship between x_i and y_i for *Ce*, *Oe*, and *DC* is such that the precision for an optimal allocation based on $\sqrt{\overline{BA}_h}$ (Eq. (14)) will be better than for an optimal allocation based on $\overline{BA}_h$ (Eq. (15)).

For each allocation method, relative precision for each accuracy estimate is computed comparing the standard error when selecting a specific per-year-biome *BA* threshold (denoted as *d*, for "dynamic" threshold) versus the alternative of selecting a fixed global *BA* threshold (denoted as *g*, for "global" threshold). This relative precision is expressed as the ratio of the standard error for *g* to that for *d* ($S_g(\hat{R})/S_d(\hat{R})$).

The precision gains of each allocation method ($n_h \propto N_h \sqrt{\overline{BA}_h}$ and $n_h \propto N_h \overline{BA}_h$) can also be compared to the precision of accuracy estimates that would have been obtained under simple random sampling. Precision comparisons were also done for two other sampling designs evaluated by Boschetti et al. (2016) with strata constructed using a fixed global threshold $p = 0.2$ of the cumulative distribution of active fires counts (MODIS MOD14A1 and MYD14A1 (Giglio et al., 2003)). These other two stratified options were stratified sampling with proportional allocation ($n_h \propto N_h$) and stratified sampling for which the standard deviation of active fire counts was used in the optimal allocation formula ($n_h \propto N_h S_{AFh}$).

Table 6
Ratio of the standard errors (SE) of Commission error ratio (*Ce*) obtained on each year and biome combination by sampling design with $n_h \propto N_h \overline{BA}_h$ against the SE obtained from design with $n_h \propto N_h \sqrt{\overline{BA}_h}$, in both cases the stratification based on optimal threshold p^*.

	2005	2006	2007	2008	2009	2010	2011
Tropical and Subtropical savanna	1.17	1.29	1.47	1.30	1.51	1.33	1.49
Tropical Forest	2.00	1.44	1.55	1.90	1.58	1.89	1.59
Temperate grassland and savanna	1.73	1.71	1.83	2.30	1.90	1.76	1.83
Boreal Forest	3.83	2.95	3.10	2.77	3.52	3.21	2.75
Temperate Forest	1.56	1.60	2.44	1.87	0.99	1.72	1.68
Mediterranean Forest	3.56	2.20	3.32	3.44	2.39	3.20	1.89
Others	3.34	3.21	2.92	2.50	2.81	2.95	2.78

Table 8
As in Table 6 but for burned area in the reference data (*BAref*).

	2005	2006	2007	2008	2009	2010	2011
Tropical and Subtropical savanna	0.96	1.19	1.09	1.35	1.17	1.10	1.10
Tropical Forest	1.78	0.99	1.28	1.45	1.33	1.46	1.45
Temperate grassland and savanna	1.94	2.03	1.84	1.79	1.82	1.80	1.51
Boreal Forest	3.95	2.22	2.72	2.31	2.49	3.19	1.85
Temperate Forest	1.83	2.26	2.26	2.19	1.41	1.99	1.87
Mediterranean Forest	3.61	2.26	3.51	3.73	2.89	3.38	2.71
Others	1.67	2.23	1.26	1.84	2.13	1.79	1.93

Table 10
As in Table 9 but for omission error ratio (*Oe*).

	2005	2006	2007	2008	2009	2010	2011
Tropical and Subtropical savanna	0.99	0.99	0.99	1.01	0.99	0.97	0.97
Tropical Forest	1.01	0.97	0.94	0.96	1.00	1.00	0.97
Temperate grassland and savanna	1.00	1.00	1.00	1.00	1.01	1.00	1.01
Boreal Forest	1.10	1.04	1.37	1.03	2.02	0.97	1.05
Temperate Forest	0.51	1.02	0.98	1.01	0.60	0.99	0.99
Mediterranean Forest	0.90	1.04	0.98	0.95	0.98	0.98	1.01
Others	1.00	1.11	1.03	0.89	0.99	1.06	1.03

2.6. Illustration of the validity of confidence intervals

As is seen in the previous sections precision of accuracy estimates are expressed with a standard error. Therefore, the validity of a confidence interval depends on the normality of the expected sampling distribution of the accuracy estimate. Although the per-sampling unit accuracy estimates may follow a strongly skewed distribution (Padilla et al., 2014a), probability sampling theory specifies that the sampling distribution of an estimator will tend to approach normality as sample size increases (Cochran, 1977; Section 2.15).

The objective of this illustration is to demonstrate the validity of the normal approximation and confidence interval coverage properties of intervals produced using the stratified sampling design and sample size available for the Fire_cci project ($n = 100$ per year). The confirmation study uses the hypothetical population mentioned in the previous Section. As in Stehman (1997), sample of units were selected from a hypothetical population large number of times, 1000 times in this study. Each sample is selected following the sampling designs presented in the current article, i.e. using the stratification and sample allocation methods presented in the previous sections.

3. Results

3.1. Stratification and sample allocation

Using the yearly sample size of 100 (as determined by the foreseeable budget of the Fire_cci project), the sample size allocation for each year and biome was established using Eqs. (14) and (15). Then, each year-biome was divided into two parts with an optimal *BA* threshold (p^*) selected to minimize $V(\widehat{BA})$. Two examples of the threshold selection for allocation $n_h \propto N_h\sqrt{\overline{BA}_h}$ are provided in Fig. 5, one for Tropical and Subtropical savanna and the other for Temperate grassland and savannah, both for year 2003. The maximum $V(\widehat{BA})$ on a year-biome occurred when it was not divided, threshold p at 1 and 0, and $V(\widehat{BA})$ reached the minimum at approximately $p = 0.2$. Fig. 6 shows two examples for the two allocations presented ($n_h \propto N_h\sqrt{\overline{BA}_h}$ and $n_h \propto N_h\overline{BA}_h$) on a common year-biome and illustrates how minimum $V(\widehat{BA})$ occurred at different p and hence how the optimal stratification can vary with the allocation used.

Tables 2 and 3 show the optimal thresholds p^* for all year-biome strata. For the full set of year by biome strata, the 25th and 75th percentiles of p^* were 0.18 and 0.29 for allocation $n_h \propto N_h\sqrt{\overline{BA}_h}$ and 0.14 and 0.21 for allocation $n_h \propto N_h\overline{BA}_h$. The large differences in the optimal allocation determined from the two methods (Tables 2–5) translated to different spatial distributions of the sample units (Figs. 7 and 8), more homogenous in allocation $n_h \propto N_h\sqrt{\overline{BA}_h}$ than in $n_h \propto N_h\overline{BA}_h$.

3.2. Precision comparison for global accuracy estimates

The relative precision achieved by selecting *BA* thresholds specific for each year-biome (named as method *d* for "dynamic" thresholds) to selecting a fixed global *BA* threshold (named method *g* for "global" threshold) tended to be a minimum at a global threshold of $p = 0.2$ for the two allocation methods $n_h \propto N_h\sqrt{\overline{BA}_h}$ and $n_h \propto N_h\overline{BA}_h$ (Fig. 9). The relative precision was close to 1 for a the range of values of p between 0.1 and 0.4 indicating a "flat optimum" in terms of choice of p. The gain in precision from using a dynamic threshold was remarkably small for allocation $n_h \propto N_h\overline{BA}_h$ for estimating *Ce* and *relB*.

Fig. 10 shows the relative precision of the four sampling designs analysed to the precision obtained under a simple random

Table 9
Ratio of the standard errors (SE) of Commission error ratio (*Ce*) obtained on each year and biome combination by sampling design with $n_h \propto N_h\sqrt{\overline{BA}_h}$ and stratification based on fixed threshold of $p = 0.2$ against the SE obtained from design with $n_h \propto N_h\sqrt{\overline{BA}_h}$ and optimal threshold p^*.

	2005	2006	2007	2008	2009	2010	2011
Tropical and Subtropical savanna	0.90	0.99	0.99	0.96	0.94	0.93	0.99
Tropical Forest	1.01	0.98	0.97	1.01	0.99	1.02	0.93
Temperate grassland and savanna	1.00	1.00	1.01	1.00	0.91	1.00	1.00
Boreal Forest	1.58	1.02	1.14	0.97	2.01	0.96	1.01
Temperate Forest	0.74	1.03	1.02	0.98	0.73	0.98	0.97
Mediterranean Forest	0.95	1.03	0.99	0.95	0.98	0.93	1.00
Others	1.07	1.00	1.00	1.00	0.99	0.99	1.32

Table 11
As in Table 9 but for burned area in the reference data (*BAref*).

	2005	2006	2007	2008	2009	2010	2011
Tropical and Subtropical savanna	0.98	0.98	0.98	1.01	0.98	0.97	0.96
Tropical Forest	0.97	0.91	0.87	0.92	0.99	1.00	0.99
Temperate grassland and savanna	1.01	1.00	0.97	1.00	1.01	1.00	1.01
Boreal Forest	0.96	1.00	1.12	0.99	1.79	0.96	1.01
Temperate Forest	0.96	1.01	0.98	1.02	0.96	1.01	1.01
Mediterranean Forest	0.89	1.04	0.98	0.94	0.98	0.98	1.01
Others	0.99	1.05	1.01	0.93	0.99	1.11	1.00

Table 13
As in Table 12 but for omission error ratio (*Oe*).

	2005	2006	2007	2008	2009	2010	2011
Tropical and Subtropical savanna	1.93	1.80	1.88	1.68	1.92	1.89	1.81
Tropical Forest	2.18	2.54	2.39	1.84	2.43	2.12	2.09
Temperate grassland and savanna	2.92	2.48	2.53	2.60	2.88	2.37	2.98
Boreal Forest	2.79	2.18	1.98	2.94	4.96	2.67	2.29
Temperate Forest	1.23	2.69	1.48	2.13	1.37	1.46	1.50
Mediterranean Forest	0.97	1.75	1.12	1.40	1.66	1.23	1.32
Others	2.40	3.05	3.17	2.19	3.32	3.30	3.57

sampling. Standard errors under simple random sampling were 2.5 to 8 times greater than those of the stratified sampling options (Fig. 10) indicating that the stratified designs offer high potential to reduce standard errors. Stratified sampling with $n_h \propto N_h \sqrt{\overline{BA}_h}$ and dynamic burned area thresholds and stratified sampling with $n_h \propto N_h S_{AFh}$ and global active fires (AF) $p = 0.2$ thresholds were clearly more efficient than the other options. The stratified design with $n_h \propto N_h \sqrt{\overline{BA}_h}$ yielded smaller standard errors for *DC* and *Oe* than the stratified option with $n_h \propto N_h S_{AFh}$ (4% and 10% smaller respectively) but the former option had larger standard errors for *Ce* and *relB* (5% and 4% larger respectively). The least efficient sampling design was clearly the proportional allocation stratified design with $n_h \propto N_h$ and global *BA* $p = 0.2$ threshold (relative precision around 2.5 for all accuracy estimates).

3.3. Precision comparisons for year by biome accuracy estimates

To evaluate the different stratified allocation options in greater depth, we examined the standard errors of the accuracy estimates for each biome by year strata. The precision comparisons by biome also provide insight into whether the relative performances of different stratification and allocation options vary by different scenarios of *BA*. These comparisons are conducted based on standard errors computed for a sample size of $n = 100$ for each combination of year and biome thus avoiding the influence of the $n_h \geq 2$ requirement (see Section 2.4). We limit the results shown here to standard errors of accuracy measures *Ce* and *Oe* and the standard error of estimated *BA* determined from the reference classification. The standard errors obtained from the stratification and allocation based on $n_h \propto N_h \overline{BA}_h$ were almost uniformly higher than the corresponding standard errors obtained based on $n_h \propto N_h \sqrt{\overline{BA}_h}$ (Tables 6–8). Hence of the two sample allocation rules (Eqs. (14) and (15)) suggested by Cochran (1977), the allocation based on $\sqrt{\overline{BA}_h}$ is preferred. This finding is also consistent with the results obtained at global scale (Fig. 10) and the advantage of the allocation using $\sqrt{\overline{BA}_h}$ relative to $\overline{BA}_h$ is extended to estimates of *DC* and *relB* (results not shown).

For the $n_h \propto N_h \sqrt{\overline{BA}_h}$ allocation, the standard errors using a fixed threshold of $p = 0.2$ were generally similar to the standard errors obtained by searching for the optimal threshold p^* (i.e., a “dynamic” threshold) separating low and high *BA* strata. For most year by biome strata, the ratio of the standard errors for estimating *BAref*, *Oe*, and *Ce* is close to 1 (Tables 9–11) indicating that the choice of a fixed global threshold of $p = 0.2$ will often be sufficient. In the Boreal Forest biome, the standard error ratio is sometimes close to 2 indicating that $p = 0.2$ is less favourable for this biome. The fact that the standard errors for the fixed threshold of $p = 0.2$ are similar to the standard errors for a dynamic threshold is not surprising given that the dynamic threshold chosen is often close to 0.2 (Table 2).

Lastly, a key question is whether the stratified sampling using mapped *BA* to determine the strata and sample allocation provides an advantage relative to simple random sampling (i.e., not using the *BA* map in the sampling design). The ratios of the standard errors for simple random sampling divided by the corresponding standard errors of the stratified design using $n_h \propto N_h \sqrt{\overline{BA}_h}$ are almost all >1 providing strong evidence that the stratified option enhances precision of the accuracy and *BAref* estimates (Tables 12–14). Many of the ratios exceed 1.5 indicating that the magnitude of the gain in precision is substantial (i.e., the standard error of simple random sampling is 1.5 times greater than the standard error of the stratified option). Although there are a few cases (biomes and years) for which simple random sampling has a smaller standard error than stratified (all occurred in the Mediterranean

Table 12
Ratio of the standard errors (SE) of Commission error ratio (*Ce*) obtained on each year and biome combination by simple random sampling against the SE obtained from design with $n_h \propto N_h \sqrt{\overline{BA}_h}$ and stratification based on optimal threshold p^*.

	2005	2006	2007	2008	2009	2010	2011
Tropical and Subtropical savanna	1.51	1.59	1.41	1.48	1.44	1.36	1.43
Tropical Forest	1.55	2.02	1.88	1.40	1.82	1.41	1.61
Temperate grassland and savanna	2.33	2.00	2.05	1.59	1.95	1.95	2.22
Boreal Forest	2.20	1.93	1.86	2.20	3.23	2.22	1.64
Temperate Forest	1.29	2.55	1.10	2.07	1.86	1.62	1.71
Mediterranean Forest	1.79	1.55	1.35	1.39	1.61	1.45	2.08
Others	1.12	1.90	1.76	1.36	1.98	1.88	2.25

Table 14
As in Table 12 but for burned area in the reference data (*BAref*).

	2005	2006	2007	2008	2009	2010	2011
Tropical and Subtropical savanna	2.41	2.30	2.24	2.10	2.47	2.27	2.20
Tropical Forest	1.81	2.44	2.10	1.89	2.39	2.16	2.26
Temperate grassland and savanna	3.36	2.60	2.50	2.79	3.06	2.18	3.28
Boreal Forest	2.17	2.37	2.36	2.98	4.54	2.92	3.13
Temperate Forest	1.83	1.84	1.43	1.68	2.36	1.48	1.37
Mediterranean Forest	0.85	1.34	0.90	0.97	1.46	1.02	1.01
Others	2.22	3.26	3.75	2.39	3.31	3.60	3.16

biome), there are so few such cases that the stratified design can be implemented with high likelihood that it will achieve better precision.

3.4. Illustration of the validity of confidence intervals

Fig. 11 shows the cumulative distributions (black lines) of accuracy expected values on the 1000 sample realizations of sampling design with allocation method $n_h \propto N_h\sqrt{\overline{X}_h}$ and per-year-biome thresholds. As anticipated, distributions of expected values tended to normality, particularly for *DC*, *Ce* and *Oe*, and the 95% confidence intervals produced from the accuracy estimates and accompanying standard errors for each sample (horizontal grey line and black dots on the extremes of intervals) contained the population observed accuracies (vertical red lines) nearly the 95% of the times, from 90% (*relB*) to 92% (*Oe*). Very similar trends of normality were observed for allocation $n_h \propto N_h\overline{BA}_h$.

4. Discussion

Several stratification and sample allocation options for defining strata and allocating the sample to strata were described and evaluated for reducing standard errors of accuracy estimates of burned area (*BA*). The premise of stratified sampling is that the map of *BA* can be used to tailor the sampling design to reduce standard errors, and a key element of stratification is deciding a threshold of *BA* to separate the sampling units (in this case defined spatially by TSAs and temporally by available Landsat imagery) into low and high *BA* strata. The approach to choosing strata and the sample allocation is strongly influenced by the information available at the planning

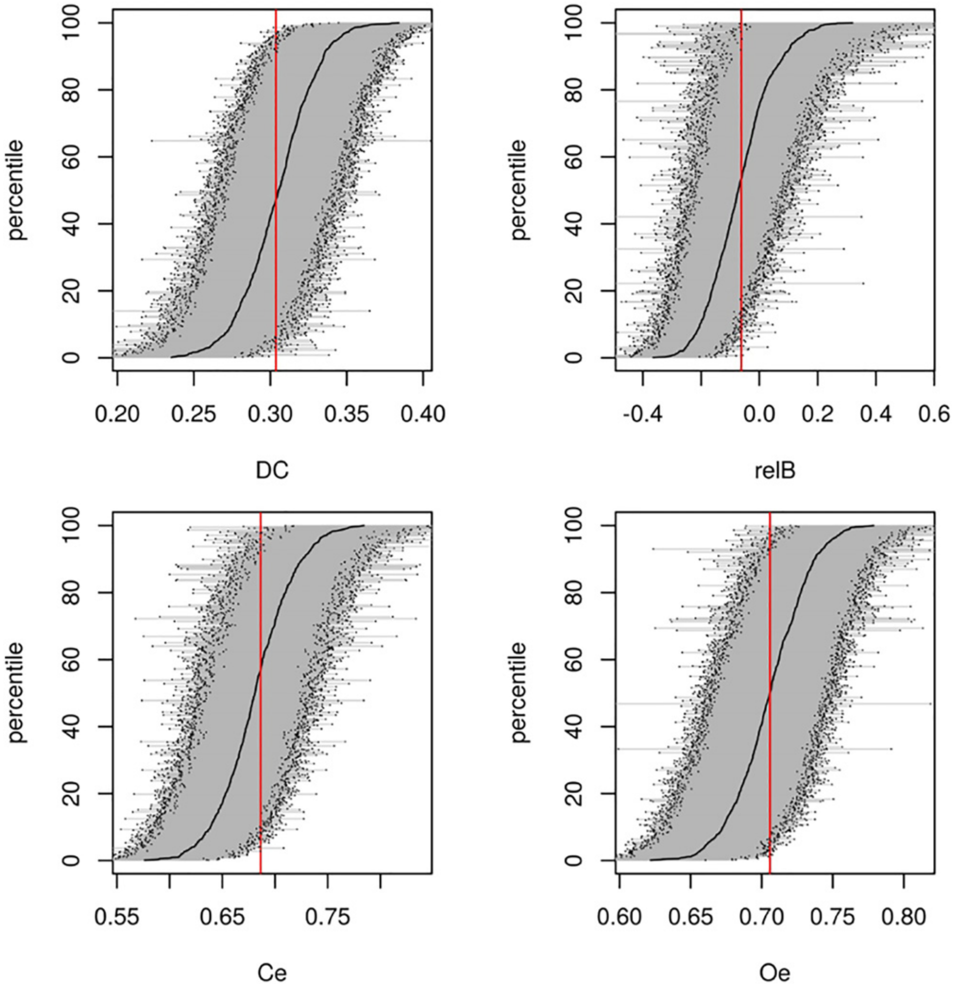

Fig. 11. Empirical cumulative distribution (black lines) of expected values for each accuracy measure on the 1000 sampling sample realizations of sampling design with allocation method $n_h \propto N_h\sqrt{\overline{X}_h}$ and per-year-biome thresholds. The 95% confidence interval of accuracy estimates on each sample is represented with a horizontal grey line and two black dots in the extremes of the intervals. Population accuracies (the target parameter) are represented with the vertical red lines.

stage of the sampling design, which is the map of *BA*. The best precision (smallest standard errors) of the accuracy estimates generally occurred at a threshold of $p = 0.2$ (i.e., the value of *BA* for the sampling units that corresponded to 20th percentile of the cumulative distribution of *BA* at each year-biome). This threshold coincided remarkably close to the threshold used by Boschetti et al. (2016). Precision at that fixed global threshold was very similar to that obtained with the stratification proposed here which allowed different per-year-biome thresholds to be chosen based on the threshold that minimized the variance of the estimated mapped *BA*, $V(\widehat{BA})$ (this approach was defined as the "dynamic" threshold). This suggests that for the particular case of inferring *BA* accuracy at global scale and with the current sampling units and biome delineations, a practical *BA* stratification would be to use fixed global $p = 0.2$ thresholds. For certain year-biome combinations, optimal thresholds were far from $p = 0.2$ (see Tables 2 and 3) so a dynamic threshold may be preferable in such cases. A benefit of the proposed per-year-biome thresholds is that this method can be applied to specific regions of interest and to other estimates of interest (e.g. accuracy of other variables).

The different optimal stratification and sample size allocations recommended by the two methods based on the stratum mean *BA* ($\overline{BA}_h$) highlights the large impacts of allocation methods on the standard errors produced by the sampling design. However two sampling designs apparently different can be similarly efficient. This is the case for one of the two designs proposed by Boschetti et al. (2016) with allocation $n_h \propto N_h S_{AFh}$ and fixed global threshold $p = 0.2$ on and active fire count distribution and the allocation $n_h \propto N_h \sqrt{\overline{BA}_h}$ with the proposed per-year-biome thresholds. Those similar efficiencies can be explained by a high correlation between active fire counts and *BA* at sampling units, leading to a proportionality between S_{AFh} and $\sqrt{\overline{BA}_h}$, and by the maximum precision found when p is around 0.2. Conversely two apparently similar sampling designs can lead to different efficiencies. The allocation $n_h \propto N_h \sqrt{\overline{BA}_h}$ can lead to much higher precision than the allocation $n_h \propto N_h \overline{BA}_h$. This is not in agreement with the findings of Hansen et al. (1946) who recommend the latter allocation specifically for populations with highly skewed distribution of x_i, with relatively few sampling units with large x_i values accounting for a large proportion of the total X. This is the case of *BA* sampling units which is expected to have highly skewed distributions across all biomes (Boschetti et al., 2016; Chou et al., 1993; Giglio et al., 2010). Conversely, the recommendation of allocation $n_h \propto N_h \sqrt{\overline{BA}_h}$ made by Cochran (1977; Section 6.14) suggests that the distribution of the ratio estimator residual variance is key to deciding which allocation method to use. In our case, the variances of y_i tend to be proportional to x_i (see Fig. 4). This relationship would in turn explain how S_{uh}^2 could tend to be proportional to $\overline{BA}_h$ and suggest using an allocation $n_h \propto N_h \sqrt{\overline{BA}_h}$. The precision evaluation using the hypothetical population provided evidence to support the recommendation to use this allocation method.

The common variations of efficiency among sampling designs observed at global scale and at year-biome level (Tables 6–8) may be caused by common distributions of the ratio estimator residual variances at the different scales. Therefore, this further emphasizes the generalization of the proposed stratification and sample allocation at any scale of interest.

The good confidence interval coverage properties obtained from the stratified design supports the validity of the confidence intervals produced from these estimates and standard errors. The frequency that 95% confidence intervals contained the true population accuracy values, close to 95%, provides assurance that even for the relatively small sample size of 100 per year the confidence intervals approximately achieved their stated nominal coverage. The coverage properties depend on the sampling distribution of the estimator and on the sample size (Cochran, 1977; Section 2.15). Hence, any increase in the sample size has a double benefit, an increase in the likelihood that confidence intervals have the stated nominal coverage, and an increase of the precision of estimates (i.e. decrease of standard errors).

As is true of any case study quantitative evaluation of precision of different sampling options, the results of this study would not necessarily generalize to all other *BA* product validation exercises. It is challenging to construct hypothetical populations that realistically mimic patterns of classification error of *BA* products. We constructed one such population, and Boschetti et al. (2016) describe another. For other populations the magnitude of the improvement in precision of accuracy estimates attributable to stratification may not match our results. However, there is reason to believe that some improvement in precision will result in most cases. Sampling theory provides support for this expected improvement as a well-chosen stratification and sample allocation almost always provides some reduction in standard errors. More importantly, we observed the reduction in standard errors due to stratification across a fairly diverse set of biomes and years suggesting that the benefit does extend to a broader set of conditions.

5. Conclusions

Several conclusions may be drawn from this study comparing standard errors of accuracy estimates for different stratification and sample allocation options. First, use of the mean mapped *BA* per stratum ($\overline{BA}_h$) to guide the definition of strata and sample allocation can produce substantial reductions (one-half to one-eighth) in standard errors relative to simple random sampling. Second, for the case study hypothetical population used in our comparative analyses, a sample size allocation proportional to $N_h \sqrt{\overline{BA}_h}$ was better than an allocation proportional to $N_h \overline{BA}_h$. Third, a fixed threshold common to all year-biome strata based on the 20th percentile of the cumulative distribution of mapped *BA* to define low and high *BA* strata was nearly as effective as a dynamic threshold tailored to each year-biome stratum. Lastly, the benefit of reduced standard errors achieved by the stratified design was found for both global accuracy estimates and for a majority of the biome-year accuracy estimates.

Acknowledgements

This work was developed within the Fire Disturbance project of the Climate Change Initiative (CCI) program of the European Space Agency (ESA).

Appendix A. Iteration process to allocate sample to ensure that $n \geq 4$ in each year-biome

$n_{year,biome}$ are initialized with Eqs. (14) or (15) and the iteration process consist on

- At year-biome strata with $n_{year,biome} < 4$
  - ○ $n_{year,biome} = 4$ (it is forced to be 4)
  - ○ $BA_{year,biome} = 0$ (it is forced to be zero)

- Recalculation of new n_{year} = previous n_{year} – n added in the previous step
- Recalculation of $n_{year,biome}$ not involved in first step with Eqs. (14) or (15) but with the updates of the previous steps
- If any $n_{year,biome} < 4$, repeat the iteration cycle keeping the updates
- The iteration process ends when all $n_{year,biome} \geq 4$.

Appendix B. Relationship between y_i and x_i on samples of validation data

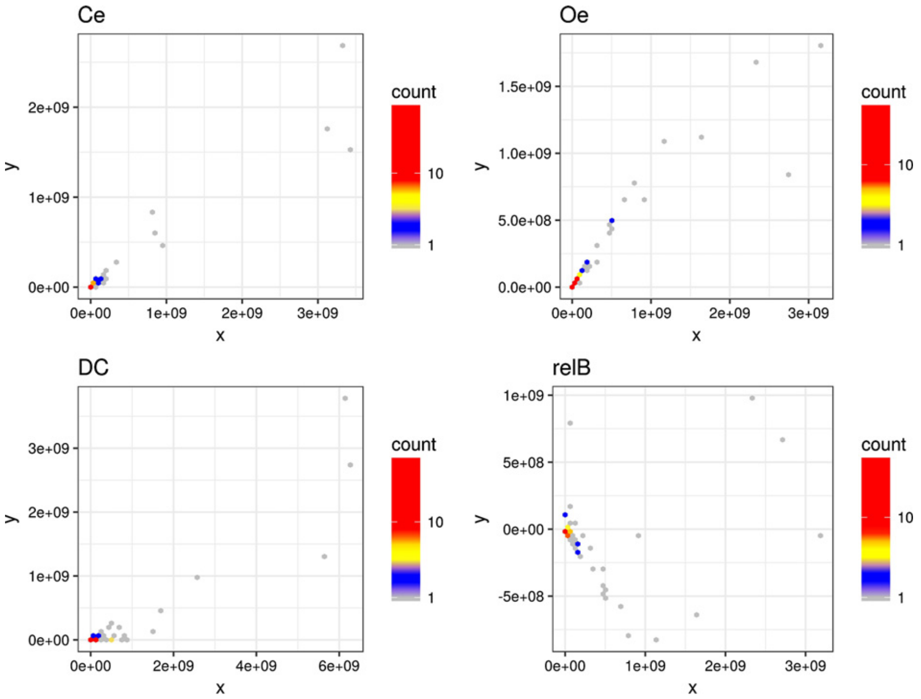

Fig. B1. Scatter plots between y_i and x_i for *Ce*, *Oe*, *DC* and *relB* on the MERIS_cci validation data presented at Padilla et al. (2015).

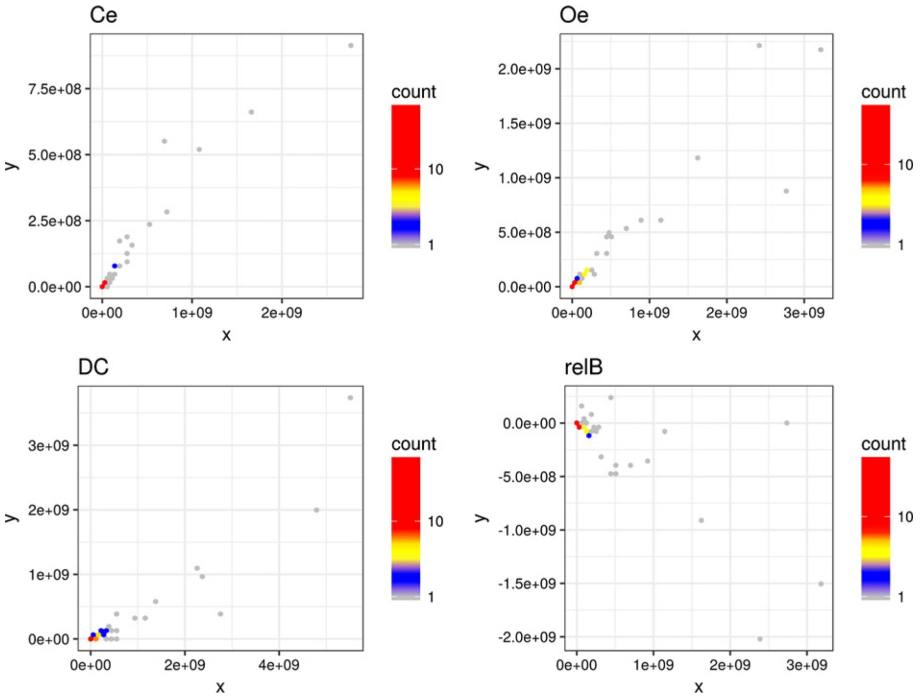

Fig. B2. As in Fig. B1 but MCD45.

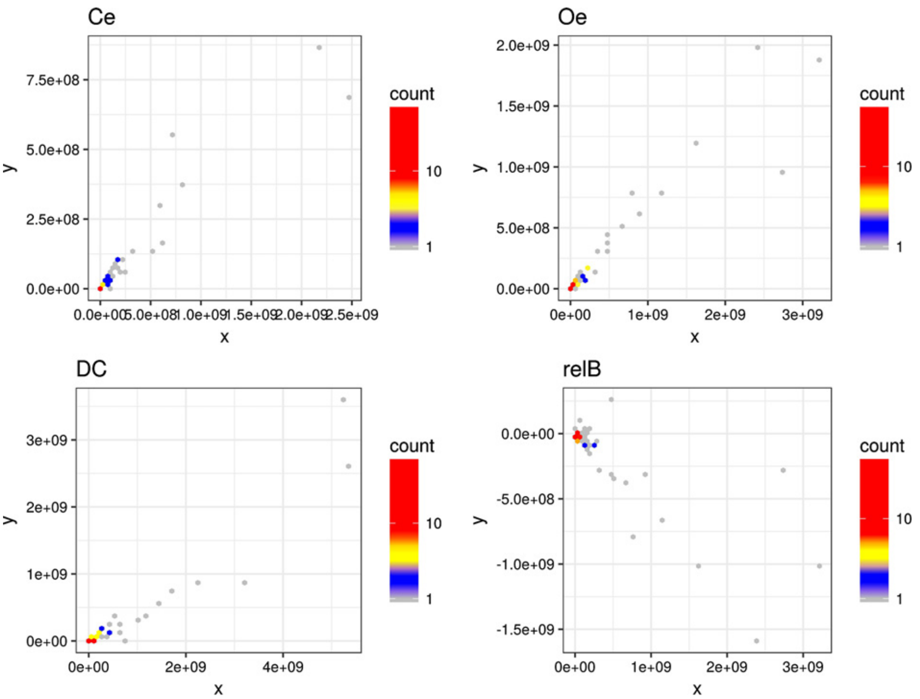

Fig. B3. As in Fig. B1 but MCD64.

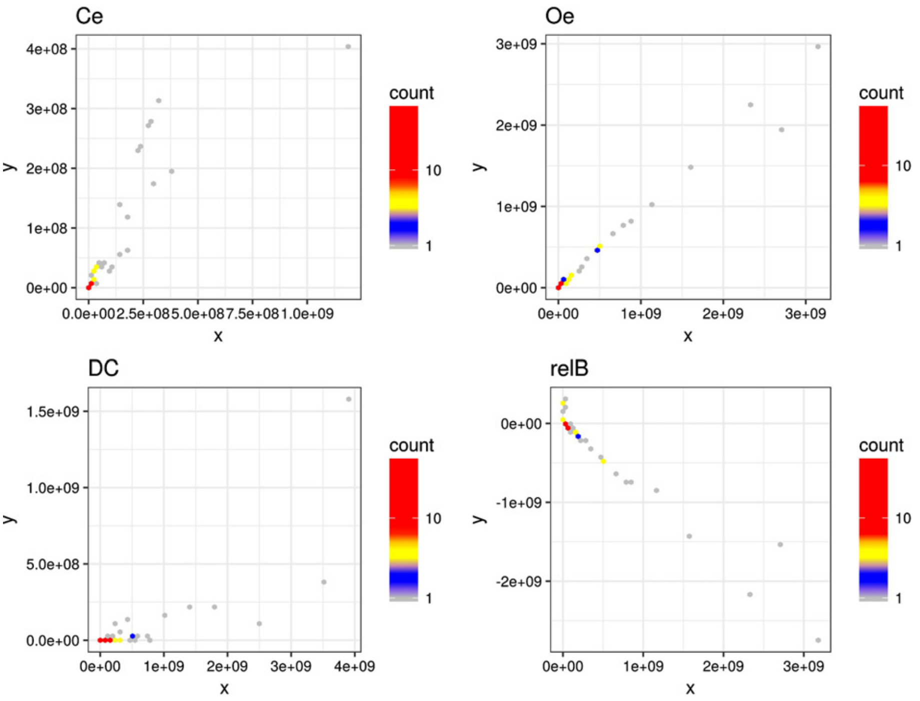

Fig. B4. As in Fig. B1 but VGT_cci.

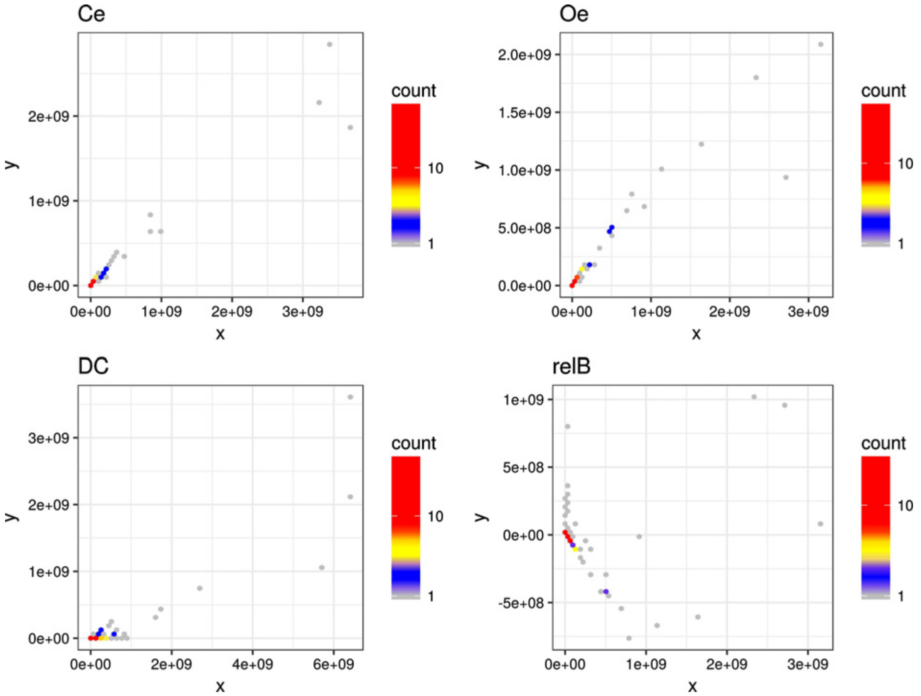

Fig. B5. As in Fig. B1 but MERGED_cci.

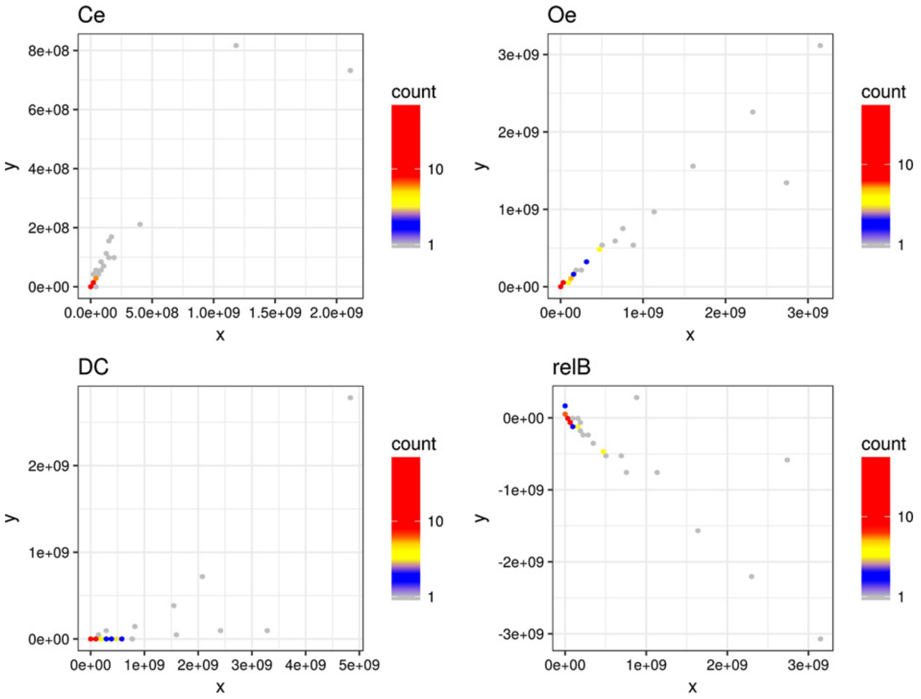

Fig. B6. As in Fig. B1 but Geoland2.

References

Bond, W.J., Keeley, J.E., 2005. Fire as a global 'herbivore': the ecology and evolution of flammable ecosystems. Trends Ecol. Evol. 20, 387–394.

Boschetti, L., Roy, D., Justice, C., 2009. In: CalVal, C. (Ed.), International Global Burned Area Satellite Product Validation Protocol. Part I – Production and Standardization of Validation Reference Data. Committee on Earth Observation Satellites, USA, pp. 1–11.

Boschetti, L., Stehman, S.V., Roy, D.P., 2016. A stratified random sampling design in space and time for regional to global scale burned area product validation. Remote Sens. Environ. 186, 465–478.

Bowman, D.M.J.S., Balch, J.K., Artaxo, P., Bond, W.J., Carlson, J.M., Cochrane, M.A., D'Antonio, C.M., DeFries, R.S., Doyle, J.C., Harrison, S.P., Johnston, F.H., Keeley, J.E., Krawchuk, M.A., Kull, C.A., Marston, J.B., Moritz, M.A., Prentice, I.C., Roos, C.I., Scott, A.C., Swetnam, T.W., Van der Werf, G.R., Pyne, S.J., 2009. Fire in the Earth system. Science 324, 481–484.

CEOS-WGCV, 2012. Working Group on Calibration and Validation - Land Product Validation Subgroup. (http://lpvs.gsfc.nasa.gov/).

Chou, Y.H., Minnich, R.A., Chase, R.A., 1993. Mapping probability of fire occurrence in San Jacinto Mountains, California, USA. Environ. Manag. 17, 129–140.

Chuvieco, E., Opazo, S., Sione, W., Del Valle, H., Anaya, J., Di Bella, C., Cruz, I., Manzo, L., López, G., Mari, N., González-Alonso, F., Morelli, F., Setzer, A., Csiszar, I., Kanpandegi, J.A., Bastarrika, A., Libonati, R., 2008. Global burned land estimation in Latin America using MODIS composite data. Ecol. Appl. 18, 64–79.

Cochran, W.G., 1977. Sampling Techniques. John Wiley & Sons, New York.

Cohen, W.B., Yang, Z., Kennedy, R.E., 2010. Detecting trends in forest disturbance and recovery using yearly Landsat time series: 2. TimeSync - tools for calibration and validation. Remote Sens. Environ. 114, 2911–2924.

Congalton, R.G., Green, K., 1999. Assessing the Accuracy of Remotely Sensed Data: Principles and Applications. Lewis Publishers, Boca Raton.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26, 297–302.

Fleiss, J.L., 1981. Statistical Methods for Rates and Proportions. John Wiley & Sons, Canada.

Forbes, A.D., 1995. Classification-algorithm evaluation: five performance measures based on confusion matrices. J. Clin. Monit. 11, 189–206.

Gallego, F.J., 2005. Stratified sampling of satellite images with a systematic grid of points. Photogramm. Eng. Remote. Sens. 59, 369–376.

GCOS, 2004. Implementation Plan for the Global Observing System for Climate in Support of the UNFCCC. World Meteorological Organization.

Giglio, L., Kendall, J.D., Mack, R., 2003. A multi-year active fire dataset for the tropics derived from the TRMM VIRS. Int. J. Remote Sens. 24, 4505–4525.

Giglio, L., Loboda, T., Roy, D.P., Quayle, B., Justice, C.O., 2009. An active-fire based burned area mapping algorithm for the MODIS sensor. Remote Sens. Environ. 113, 408–420.

Giglio, L., Randerson, J.T., van der Werf, G.R., Kasibhatla, P., Collatz, G.J., Morton, D.C., Defries, R., 2010. Assessing variability and long-term trends in burned area by merging multiple satellite fire products. Biogeosci. Discuss. 7, 1171.

Hand, D.J., 1981. Discrimination and Classification. John Wiley and Sons, New York.

Hansen, H.M., Hurwitz, W.N., Gurney, M., 1946. Problems and methods of the sample survey of business. J. Am. Stat. Assoc. 41, 173–189.

Hantson, S., Padilla, M., Corti, D., Chuvieco, E., 2013. Strengths and weaknesses of MODIS hotspots to characterize global fire occurrence. Remote Sens. Environ. 131, 152–159.

Hellden, U., 1980. A Test of Landsat-2 Imagery and Digital Data for Thematic Mapping, Illustrated by an Environmental Study in Northern Kenya. Lund University Natural Geography Institute, Sweden.

Kennedy, R.E., Yang, Z., Cohen, W.B., 2010. Detecting trends in forest disturbance and recovery using yearly Landsat time series: 1. LandTrendr - temporal segmentation algorithms. Remote Sens. Environ. 114, 2897–2910.

Latifovic, R., Olthof, I., 2004. Accuracy assessment using sub-pixel fractional error matrices of global land cover products derived from satellite data. Remote Sens. Environ. 90, 153–165.

Liu, C., Frazier, P., Kumar, L., 2007. Comparative assessment of the measures of thematic classification accuracy. Remote Sens. Environ. 107, 606–616.

Mouillot, F., Schultz, M.G., Yue, C., Cadule, P., Tansey, K., Ciais, P., Chuvieco, E., 2014. Ten years of global burned area products from spaceborne remote sensing - a review: analysis of user needs and recommendations for future developments. Int. J. Appl. Earth Obs. Geoinf. 26, 64–79.

Olson, D.M., Dinerstein, E., Wikramanayake, E.D., Burgess, N.D., Powell, G.V.N., Underwood, E.C., D'Amico, J.A., Itoua, I., Strand, H.E., Morrison, J.C., Loucks, C.J., Allnutt, T.F., Ricketts, T.H., Kura, Y., Lamoreux, J.F., Wettengel, W.W., Hedao, P., Kassem, K.R., 2001. Terrestrial ecoregions of the world: a new map of life on earth. Bioscience 51, 933–938.

Padilla, M., Stehman, S.V., Chuvieco, E., 2014a. Validation of the 2008 MODIS-MCD45 global burned area product using stratified random sampling. Remote Sens. Environ. 144, 187–196.

Padilla, M., Stehman, S.V., Litago, J., Chuvieco, E., 2014b. Assessing the temporal stability of the accuracy of a time series of burned area products. Remote Sens. 6, 2050–2068.

Padilla, M., Stehman, S.V., Ramo, R., Corti, D., Hantson, S., Oliva, P., Alonso, I., Bradley, A., Tansey, K., Mota, B., Pereira, J.M., Chuvieco, E., 2015. Comparing the accuracies of remote sensing global burned area products using stratified random sampling and estimation. Remote Sens. Environ. 160, 114–121.

Plummer, S., Arino, O., Ranera, F., Tansey, K., Chen, J., Dedieu, G., Eva, H., Piccolini, I., Leigh, R., Borstlap, G., Beusen, B., Fierens, F., Heyns, W., Benedetti, R., Lacaze, R., Garrigues, S., Quaife, T., De Kauwe, M., Quegan, S., Raupach, M., Briggs, P., Poulter, B., Bondeau, A., Rayner, P., Schultz, M., McCallum, I., 2007. An update on the GlobCarbon initiative: multi-sensor estimation of global biophysical products for global terrestrial carbon studies. Envisat Symposium 2007. Montreux, Switzerland.

Randerson, J.T., Chen, Y., van der Werf, G.R., Rogers, B.M., Morton, D.C., 2012. Global burned area and biomass burning emissions from small fires. J. Geophys. Res. 117.

Roy, D.P., Boschetti, L., 2009. Southern Africa validation of the MODIS, L3JRC, and GlobCarbon burned-area products. IEEE Trans. Geosci. Remote Sens. 47, 1032–1044.

Roy, D.P., Boschetti, L., Justice, C.O., Ju, J., 2008. The collection 5 MODIS burned area product - global evaluation by comparison with the MODIS active fire product. Remote Sens. Environ. 112, 3690–3707.

Stehman, S.V., 1997. Estimating standard errors of accuracy assessment statistics under cluster sampling. Remote Sens. Environ. 60, 258–269.

Tansey, K., Grégoire, J.-M., Defourny, P., Leigh, R., Pekel, J.-F., Bogaert, E., Bartholome, E., 2008. A new, global, multi-annual (2000–2007) burnt area product at 1 km resolution. Geophys. Res. Lett. 35, L01401. http://dx.doi.org/10.1029/2007GL03156.

van der Werf, G.R., Randerson, J.T., Collatz, G.J., Giglio, L., Kasibhatla, P.S., Arellano, A.F., Olsen, S.C., Kasischke, E.S., 2004. Continental scale-partitioning of fire emissions during the 1997 to 2001 El Niño/La Niña period. Science 303, 73–76.

van der Werf, G.R., Randerson, J.T., Giglio, L., Collatz, G.J., Mu, M., Kasibhatla, P.S., Morton, D.C., DeFries, R.S., Jin, Y., van Leeuwen, T.T., 2010. Global fire emissions and the contribution of deforestation, savanna, forest, agricultural, and peat fires (1997–2009). Atmos. Chem. Phys. 10, 11707–11735.