Supplement of Biogeosciences, 17, 1673–1683, 2020
https://doi.org/10.5194/bg-17-1673-2020-supplement

Biogeosciences Open Access EGU

Supplement of

An analysis of forest biomass sampling strategies across scales

Jessica Hetzer et al.

Correspondence to: Jessica Hetzer (jessica.hetzer@ufz.de)

S1 Theory of random sampling

Random sampling can be covered analytically by the central limit theorem, which states that whenever independent random variables are added, their sum converges toward a normal distribution irrespective of the original distribution. Suppose that $X_1, X_2, \ldots X_n$ is a sequence of independent identically distributed random variables with a mean $\overline{X_n}$, finite expected value $E(X_i) = \mu$, and variance $Var(X_i) = \sigma^2$. For $n \to \infty$, the distribution function, Z_n, converges to the standardized normal distribution.

$$Z_n = \frac{\sqrt{n}(\overline{X_n} - \mu)}{\sigma}$$

The probability, P_n, that $\overline{X_n} \in [0.9 \cdot \mu, 1.1 \cdot \mu]$ can be calculated as the "sample reliability", where Φ is the probability of the normal distribution, can be calculated as follows:

$$P_n = P\left(|z_n| < 0.1\sqrt{n}\frac{\mu}{\sigma}\right) = \Phi\left(0.1\sqrt{n}\frac{\mu}{\sigma}\right) - \Phi\left(-0.1\sqrt{n}\frac{\mu}{\sigma}\right)$$

$$= 2 * \Phi\left(0.1\sqrt{n}\frac{\mu}{\sigma}\right) - 1 \qquad (1)$$

The "minimum sample size", n_{min}, defines the number of samples needed under the condition that the mean of the samples does not deviate more than 10 % from the real mean biomass with a probability of at least 90 %. We determined n_{min} using Eq. (1) and the quantile of the standardized normal function, q:

$$2 \cdot \Phi\left(0.1\sqrt{n}\frac{\mu}{\sigma}\right) - 1 \geq 0.9$$

thus

$$\Phi\left(0.1\sqrt{n}\frac{\mu}{\sigma}\right) \geq \frac{1.9}{2}$$

Therefore

$$0.1\sqrt{n}\frac{\mu}{\sigma} \geq q\left(\frac{1.9}{2}\right)$$

Finally,

$$n \geq \frac{(q(0.95)\cdot\sigma)^2}{(0.1\,\mu)^2}$$

The minimum sample size here is assumed to be the minimum number of samples for which this equation is still valid.

$$n_{min} := \left\lceil \frac{(q(0.95) \cdot \sigma)^2}{(0.1\,\mu)^2} \right\rceil$$

n_{min} depends only on the mean biomass, μ, and its standard derivation, σ. In this study, this term is simplified using the coefficient of variation ($CV = \frac{\sigma}{\mu}$) as follows:

$$n_{min} = \left\lceil \frac{q(0.95)^2}{0.01} \cdot CV^2 \right\rceil \qquad (3).$$

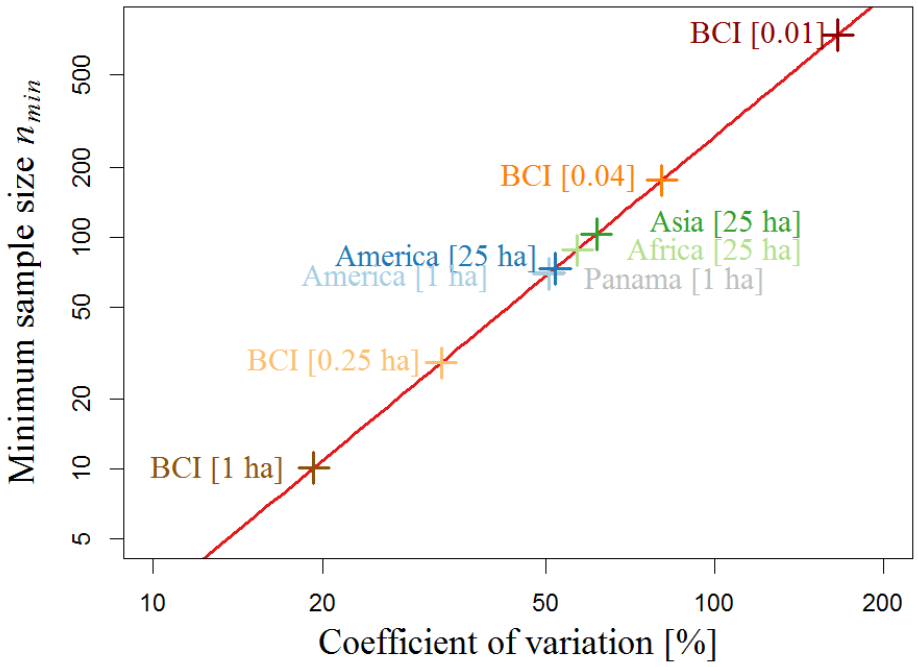

Figure S1 Relation between the minimum sample size and the coefficient of variation (CV) of the biomass distribution. The results for random sampling of Barro Colorado Island (BCI, 50 ha), Panama (~50000 km^2), America, Africa and Asia (~3-11 Mio km^2) are shown. The analyzed biomass maps are displayed with colored crosses (with the name and plot size shown in brackets). The analytical relationship between CV and the minimum sample size has been derived from equation 3 (red line). Please note that both axes are logarithmized.

Table S1 Analyzed forest biomass maps for the tropics and their minimum sample size. Shown are the forest biomass maps for **(a)** South America, Africa and Southeast Asia (Baccini et al., 2012) and for **(b)** South America, Africa and Asia & Australia (Saatchi et al., 2011) and their random sampling performance. The minimum sample size refers to the minimum number of plots to accurately estimate the mean biomass of the forest (the mean of the samples does not deviate more than 10 % from the real mean biomass with a probability of at least 90 %). The last column shows the necessary sampling area $\boldsymbol{a_{min} = A_{plot} \cdot n_{min}}$.

Map (Resolution)	Map size [Mio. km^2]	Plot size A_{plot}[ha]	CV	Minimum sample size $\boldsymbol{n_{min}}$ [plots]	Minimum total area of samples a_{min} [ha]
a)					
South America (500 m)	11.3	25	51.98	74	1850
Africa (500 m)	3.4	25	56.94	88	2200
Southeast Asia (500 m)	6.1	25	61.61	103	2575
b)					
South America (1000 m)	11.4	100	52.51	75	75000
Africa (1000 m)	3.8	100	62.23	105	105000
Asia & Australia (1000 m)	10.7	100	70.65	136	136000

S2 Point pattern summary functions for the clustered sampling approach

For the reconstructions of the clustered locations of the inventory plots, we use several point pattern summary functions that quantify the spatial structure of the pattern within distances of up to 100 km: (I) the probabilities $p(k, r)$ that the typical point of the pattern has k neighbors within distance interval r – bin/2, r + bin/2, where bin = 500 m is the resolution of the map and r = 0.5, 2.5, 7.5, 17.5, 25, and 50 km. (II) the distribution function $D(r)$ of the distances to the nearest neighbor; (III), the average number of points at distance r from the points of the pattern given by $\lambda(2\pi r)g(r)$ where λ is the density of the pattern and $g(r)$ the pair correlation function; (IV) the average number of points at a distance r from the points of the pattern given by $\lambda K(r)$, where $K(r)$ is Ripley's K; (V) $H_s(r)$ the spherical contact distribution; (VI) the distribution functions $D_k(r)$ of the k'th nearest neighbor. For further details see Wiegand, He and Hubbell, (2013).

S3 Downscaling of the South America map

To downscale the 500 m resolution map of South America to a 100 m resolution, we used statistical relationships derived from the Panama map at 100 m and the 500 m resolution. Therefore, 25 plots from the original map (100 m resolution) were aggregated to a mean value. The standard deviation of those 25 plots at a 100 m resolution sd_{100} can then be plotted against the mean value of the plots, which can be interpreted as the aggregated value at a 500 m scale AGB_{500} (see Fig. S2 a).

For the first downscaling strategy (D1) we transferred the derived relationships to the South America map. After creating classes over the AGB of 1 t/ha, each AGB value of 25 ha plot of the South American map was assigned to 25 plots of 1 ha drawing random values from a normal distribution $N(AGB_{500}, (sd_{100})^2)$. If the South American plot had an AGB value higher than the maximum value of Panama, the created plots were drawn from a normal distribution with the standard deviation of the maximum class. Negative biomass plots were set to zero.

We analyzed also a second downscaling strategy (D2). Here we assume that the variation of subplots increases linear with biomass (compare Figure S1). For this down-scaling strategy the linear trend resulting from AGB values smaller than 100 t/ha is continued for larger biomass values. Each AGB value of the South American map (1 pixel = 25ha) was assigned to 25 plots of 1 ha drawing random values from a normal distribution $N(AGB_{500}, (m*AGB_{500}+t)^2)$ with slope m and intercept t as coefficients of the linear regression.

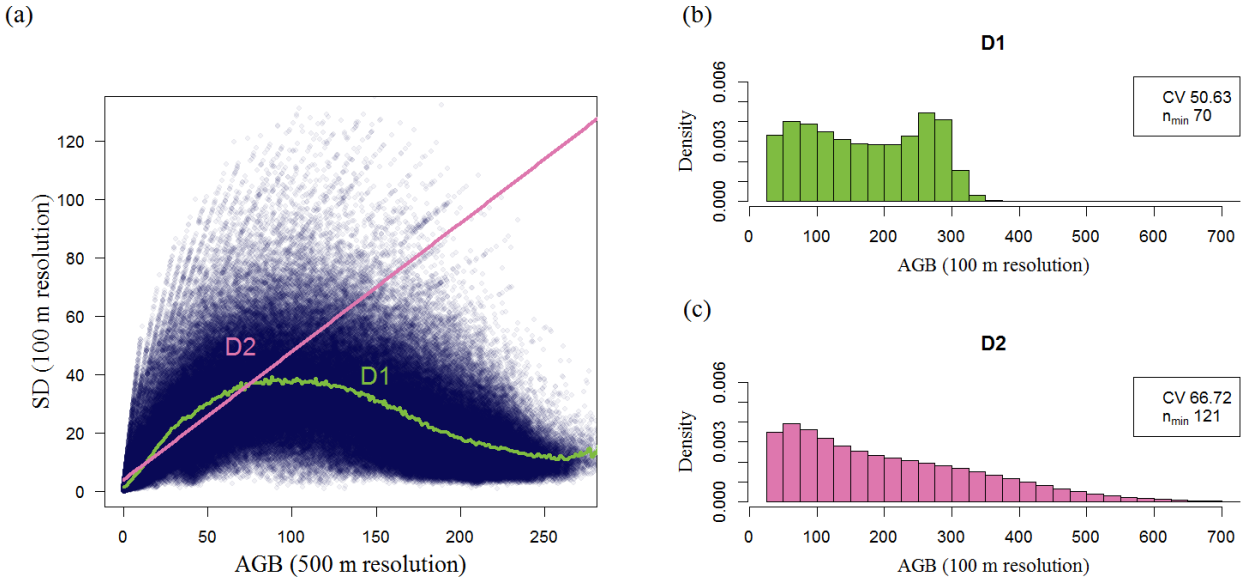

Figure S2 Comparison of downscaling approaches. **(a)** Subplot heterogeneity in the Panama biomass map (500 m resolution). Shown is the aboveground biomass (AGB) of plots at a 500 m resolution (x-axis) and the standard deviation (SD) of the associated 25 subplots at a 100 m resolution (y-axis). Each dot represents one plot from the Panama map (~300,000 plots).The green line represents the downscaling approach D1, as it was implemented in the current study (Table 1). The second downscaling approach D2 shown in pink is based on an increasing linear relationship. **(b)-(c)** Aboveground biomass distribution of South America at a 100 m resolution for the two analysed downscaling approaches. Coefficient of

variation (CV) and the minimum sample size (n_{min}) of randomly chosen 1 ha plots are shown at the upper right corner for each biomass distribution.

S4 Spatial clustering of the biomass

To quantify spatial correlation structures in the biomass map we used the mark variogram $\gamma_{mm}(r)$ (Illian et al. 2008; Wiegand and Moloney 2014). Each grid cell of the map has x-y coordinates and the associated biomass value is the mark. We then consider all pairs of cells that are distance r apart and determined the mean value of the test function $t_4(m_1, m_2) = (m_1 - m_2)^2/2$ where m_1 is the biomass mark of the first cell and m_2 that of the second cell. The mark variogram has small values if the biomasses of cells that are distance r apart are in general similar to each other and large values if the biomasses are dissimilar. To test for significant spatial structure we compared the observed mark variogram to that of a null model where we randomly permutated the biomass values among the all forested cells.

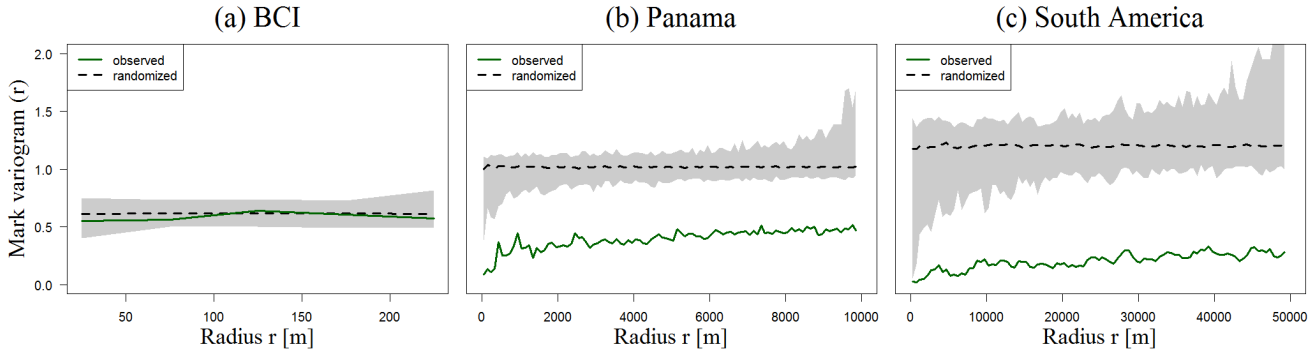

Figure S3 Spatial structure of the biomass maps. Shown is the observed mark variogram (green line, for details see Illian *et al.* (2008)) for maps of Barro Colorado Island (50 m resolution, 50 ha), Panama (100 m resolution, ~50000 km^2) and South America (500 m resolution, 15 Mio km^2). Values of the dashed black line show the expectation under the null model of random distribution of biomass (gray color displays the 99% simulation envelopes). Values below the envelopes indicate clustering (more similar biomass values at short distances than expected by the null model), values above indicate overdispersion (more dissimilar biomass values than expected by the null model, e.g., comparable to a chessboard). For further details on the method, see Wiegand and Moloney (2014). Please note that for Panama and America, we use subsets of 10,000 plots. **(a)** No significant spatial structure was observed for the 50m resolution for Barro Colorado Island (BCI)**. (b)-(c)** In contrast to BCI, the observed biomasses for Panama and America are below simulation envelopes, indicating biomass values are within distances of 10 km and 50 km more similar to each other than expected by a random distribution of biomasses over all forested cells.

S5 Transect sampling in South America

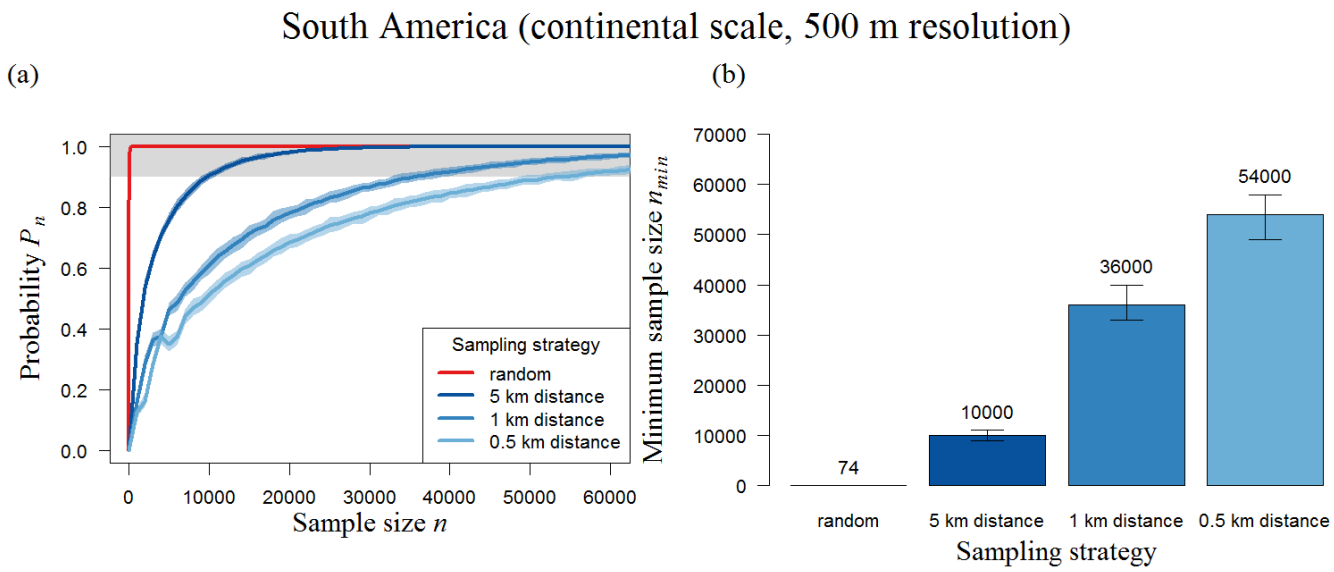

Figure S4 Results of transect sampling for different biomes of South America. Left: Simulation results showing the number of plots and probability ($\boldsymbol{P_n}$) that the mean biomass of those plots reflects the mean biomass of the whole forest (for details, see Methods). We focus on three strategies using distances of 500 m, 1 km and 5 km between plots (shown in blue) and compare them to random sampling (red). The area around each line represents the 95 % confidence intervals derived from 100 repetitions (total of 1000*100 runs for each sample size). The upper boundary (gray) marks sample sizes with accurate biomass estimations ($\boldsymbol{P_n \geq 90}$ %). Right: Necessary number ($\boldsymbol{n_{min}}$) of 1 ha plots for Panama and of 25 ha plots for South America (error bars show the 95 % confidence intervals of 100 repetitions).

References

Baccini, A., Goetz, S. J., Walker, W. S., Laporte, N. T., Sun, M., Sulla-Menashe, D., Hackler, J., Beck, P. S. A., Dubayah, R., Friedl, M. A., Samanta, S. and Houghton, R. A.: Estimated carbon dioxide emissions from tropical deforestation improved by carbon-density maps, Nat. Clim. Chang., 2(3), 182–185 [online] Available from: http://dx.doi.org/10.1038/nclimate1354, 2012.

Illian, J. B., Penttinen, A., Stoyan, H. and Stoyan, D.: Statistical Analysis and Modelling of Spatial Point Patterns, John Wiley and Sons., 2008.

Saatchi, S. S., Harris, N. L., Brown, S., Lefsky, M., Mitchard, E. T. A. and Salas, W.: Benchmark map of forest carbon stocks in tropical regions across three continents, , 108(24), doi:10.1073/pnas.1019576108, 2011.

Wiegand, T. and Moloney, K.: A Handbook of Spatial Point Pattern Analysis in Ecology., 2014.

Wiegand, T., He, F. and Hubbell, S. P.: A systematic comparison of summary characteristics for quantifying point patterns in ecology, Ecography (Cop.)., 36(1), 092–103, doi:10.1111/j.1600-0587.2012.07361.x, 2013.