


MODIS VCF should not be used to detect discontinuities in tree cover due to binning bias. A comment on Hanan et al. (2014) and Staver and Hansen (2015)

France Gerard¹  | Danny Hooftman^{1,2} | Frank van Langevelde³ |
Elmar Veenendaal⁴ | Steven M. White^{1,5} | Jon Lloyd^{6,7}

¹Centre for Ecology and Hydrology
Wallingford, Crowmarsh Gifford, OX10 8BB,
United Kingdom

²Lactuca: Environmental Data Analyses and
Modelling, Diemen, 1112NC,
The Netherlands

³Resource Ecology group, Wageningen
University, Wageningen, The Netherlands

⁴Plant Ecology and Nature Conservation,
Wageningen University, Wageningen,
The Netherlands

⁵Wolfson Centre for Mathematical Biology,
Mathematical Institute, Radcliffe
Observatory Quarter, Oxfordshire,
United Kingdom

⁶Department of Life Sciences, Faculty of
Natural Sciences, Imperial College (Silwood
Park), Ascot, United Kingdom

⁷Centre for Tropical Environment and
Sustainability Sciences (TESS) and College of
Marine and Environmental Sciences, James
Cook University, Cairns, Queensland,
Australia

Correspondence

France Gerard, Centre for Ecology and
Hydrology Wallingford, Crowmarsh Gifford,
OX10 8BB, United Kingdom.

Email: ffg@ceh.ac.uk

Editor: Thomas Gillespie

Funding information

NERC, Grant Number: NE/D005469/1

Abstract

In their recent paper, Staver and Hansen (*Global Ecology and Biogeography*, 2015, 24, 985–987) refute the case made by Hanan et al. (*Global Ecology and Biogeography*, 2014, 23, 259–263) that the use of classification and regression trees (CARTs) to predict tree cover from remotely sensed imagery (MODIS VCF) inherently introduces biases, thus making the resulting tree cover unsuitable for showing alternative stable states through tree cover frequency distribution analyses. Here we provide a new and equally fundamental argument for why the published frequency distributions should not be used for such purposes. We show that the practice of pre-average binning of tree cover values used to derive cover values to train the CART model will also introduce errors in the frequency distributions of the final product. We demonstrate that the frequency minima found at tree covers of 8–18%, 33–45% and 55–75% can be attributed to numerical biases introduced when training samples are derived from landscapes containing asymmetric tree cover distributions and/or a tree cover gradient. So it is highly likely that the CART, used to produce MODIS VCF, delivers tree cover frequency distributions that do not reflect the real world situation.

KEYWORDS

alternative stable states, forest, frequency distribution, MODIS VCF, remote sensing, savanna, tree cover

The MODIS VCF tree cover product of Hansen, DeFries, Townshend, Marufu, and Sohlberg (2002) provides world-wide estimates of percentage tree cover derived from MODIS data. Discontinuities in tree cover frequency distributions derived from these data have been used to support the hypothesis that the observed distributions of forest, savanna and grassland vegetation in the tropical and boreal regions of the world represent alternative stable states for equivalent environmental conditions (Favier et al., 2012; Hirota, Holmgren, Van Nes, & Scheffer, 2011; Murphy & Bowman, 2012; Scheffer, Hirota, Holmgren,

Van Nes, & Chapin, 2012; Staver, Archibald, & Levin, 2011; Xu et al., 2016). However, recently Hanan, Tredennick, Prihodko, Bucini, and Dohn (2014, 2015) suggested that the adopted classification and regression trees (CART) approach used to produce the MODIS VCF tree cover estimates introduced a systematic bias which makes the MODIS VCF product inappropriate for the analysis of percentage tree cover frequency distributions. This point was countered by Staver and Hansen (2015), who argued (i) that the approach taken by Hanan et al. (2014), using simulated EO data and pseudosatellite metrics to

demonstrate that an artificial bias is generated by the CART approach, does not reflect the complexity and variability of landscapes and vegetation across the globe and (ii) that the CART model used by Hanan et al. (2014) was highly pruned with very few nodes (9 nodes) compared with that used for the MODIS VCF product (109 nodes) resulting in a less smooth gradient of percentage tree cover values. In this analysis we show that the VCF tree cover product is likely to also contain systematic bias which is introduced in the pre-processing of calibration data used to train the CART model. Such errors would exacerbate the problem of aggregation in the CART predictions already demonstrated by Hanan et al. (2014).

The VCF product is the result of a continuously developing method which has been documented in a succession of publications (Defries, Hansen, Townshend, Janetos, & Loveland, 2000; DeFries, Townshend, & Hansen, 1999; Hansen, Townshend, Defries, & Carroll, 2005; Hansen et al., 2003; Hansen, DeFries, Townshend, Marufu, et al., 2002; Hansen, DeFries, Townshend, Sohlberg, et al., 2002). The most recent version is referred to as MOD44B collection 5, which is available at a 250-m resolution and supersedes the previous collections, including the 500-m MOD44B collection 3. Collection 5 was used in the most recent study of Xu et al. (2016), while collection 3 was used in the majority of the publications about discontinuities in tree cover distributions (Favier et al., 2012; Hirota et al., 2011; Murphy & Bowman, 2012; Scheffer et al., 2012; Staver et al., 2011).

The approach used to create the VCF product relies on two critical components: (i) the creation of training samples and (ii) the implementation of a CART model that derives percentage tree cover from a collection of MODIS-based metrics. The design (the number of nodes, the choice of regression variables from the pool of MODIS-based metrics, and the regressions) of the CART model is determined by the training samples. In other words, during training, the design of the CART is tailored to best reproduce the percentage tree cover values of the training samples. The method for deriving the samples used to train the

CART model has remained the same between collections (DiMiceli et al., 2011; Townshend et al., 2011) and so the argument below applies to all VCF versions, including the most recent MOD44B collection 5.

Summarizing from Hansen, DeFries, Townshend, Sohlberg, et al. (2002) (also the VCF User Guide), training samples are created as follows: 30-m Landsat Thematic Mapper image pixels are assigned one of four discrete percentage tree cover classes (0%, 25%, 50%, 80%) with their boundaries defined as (0–10%, 11–40%, 41–60%, 61–100%), respectively. These classified 30-m Thematic Mapper images are then re-gridded to produce 500-m (or 250-m) training pixels, matching the size of the MODIS pixels. The resulting percentage tree cover values represent the average of the 30-m pixel class values contained within each 500-m pixel. In short, the training values represent weighted averages of the discrete values 0%, 25%, 50% and 80%, with the weights defined by the number of 30-m pixels found within the 500-m pixels having one of the respective four values. We here evaluated the impact of deriving a percentage tree cover gradient in this manner, which we will henceforth refer to as ‘averaging with pre-average binning’, by first (i) considering a range of one-dimensional cover distributions through a semi-analytical experiment and then (ii) evaluating the effect of spatial autocorrelation through a two-dimensional Monte Carlo simulation experiment.

The semi-analytical experiment involved simulations using the Beta distributions (Taboga, 2012). The Beta distribution enabled us to represent a wide variety of possible continuous tree cover distributions found within a training pixel (i.e. a 500-m pixel) by varying the shape parameters α and β of the distribution (see Figure S1.1 in the Supporting Information) and the percentage tree cover range (i.e. 0–100%, 0–80%, 20–100%) (see Appendix S1 for details). Pre-average binning into bin-classes 0.5% (c. 0%), 25%, 50% and 80% is achieved through a subdivision of the distribution $\mathbf{B}(\alpha, \beta)$ into four pieces, using three break points $z_1 = 0.1$, $z_2 = 0.4$ and $z_3 = 0.6$ and calculating the inferred mean (\hat{X}):

$$\langle \hat{X} \rangle = \frac{\int_0^{z_1} \frac{z_1}{2} \mathbf{B}(\alpha, \beta) dx + \int_{z_1}^{z_2} \left(z_1 + \frac{z_2 - z_1}{2} \right) \mathbf{B}(\alpha, \beta) dx + \int_{z_2}^{z_3} \left(z_2 + \frac{z_3 - z_2}{2} \right) \mathbf{B}(\alpha, \beta) dx + \int_{z_3}^1 \left(z_3 + \frac{1 - z_3}{2} \right) \mathbf{B}(\alpha, \beta) dx}{\int_0^1 \mathbf{B}(\alpha, \beta) dx} \quad (1)$$

where $\mathbf{B}(\alpha, \beta) = x^{\alpha-1} (1-x)^{\beta-1}$ and α and β are the shape parameters (≥ 0); the denominator $\int_0^1 \mathbf{B}(\alpha, \beta) dx$ can be regarded as a normalizing factor. Full details of the derivation of Equation 1 can be found in Appendix S1. Testing for any bias was undertaken through a comparison of the inferred mean (\hat{X}) with the true mean \bar{x} . Specifically, we tested for many random values of α and β (> 0.5 and ≤ 6) postulating that if there was no bias introduced through binning, the inferred mean would match the true mean.

Results from this semi-analytical experiment clearly demonstrate that binning percentage tree covers prior to averaging must introduce biases in the CART training samples when the original tree cover distribution within a tile is asymmetric. It is only when the distribution of percentage tree cover is perfectly symmetric that no error in the

binning averages will occur (see Equation A5 in Appendix S1). The magnitude of the bias and its location across the percentage tree cover axis is defined by the shape of the distribution and the percentage tree cover range within the tile, but a consistent tendency is for an overestimation in averages near the 20–30% and 65% cover values and underestimation in averages near the 10–15% and the 40–50% cover values (Figure 1). This pattern, although small, matches the discrepancy observed in the validation data for Africa shown in fig. 1(b) of Staver and Hansen (2015).

Monte Carlo simulations were designed to deliver theoretical two-dimensional landscapes of percentage tree cover, in the form of a 1000 × 1000 grid of values in a similar fashion to the above approach, but in this case allowing for spatial autocorrelation effects to be examined.

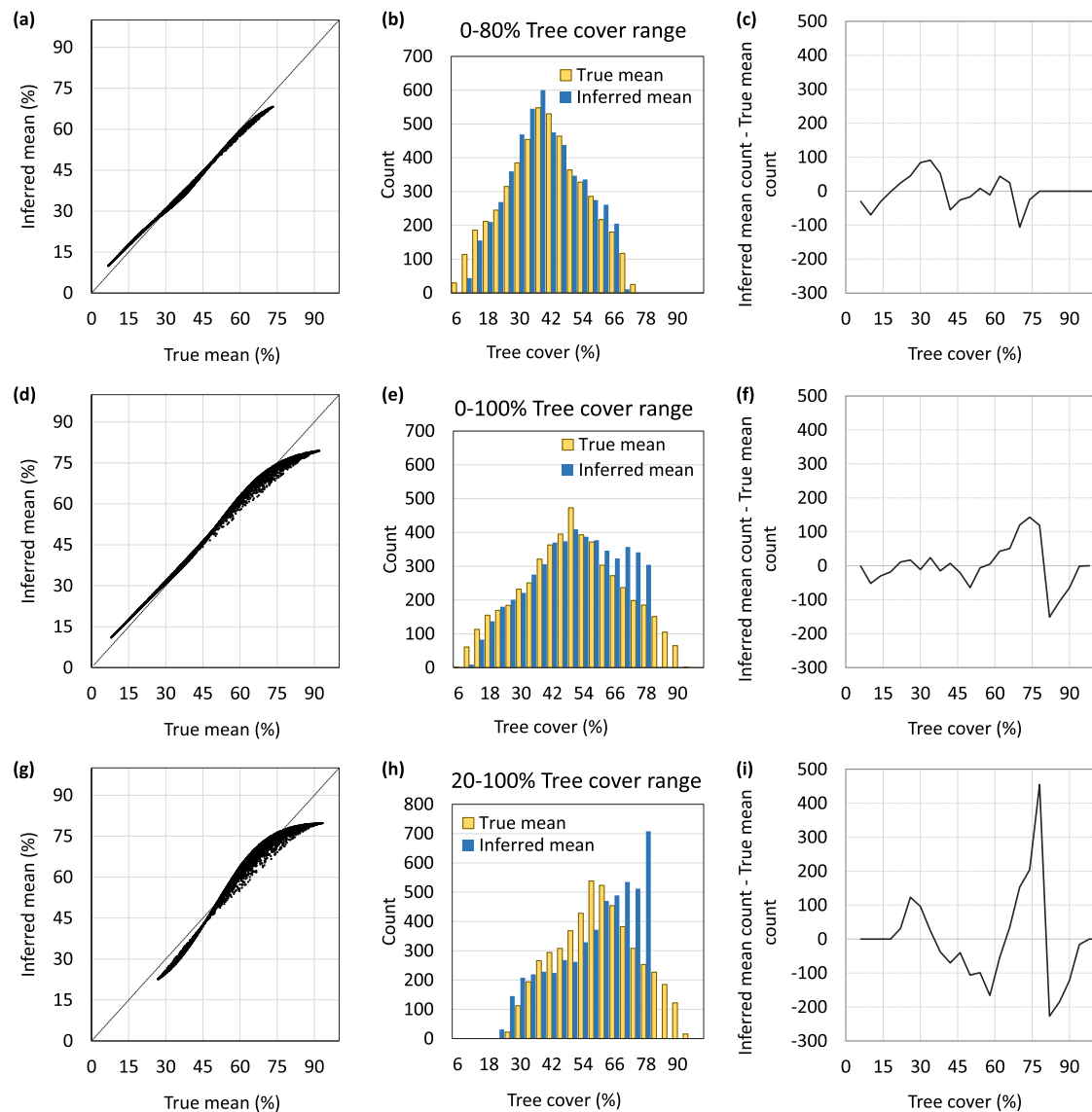


FIGURE 1 (a), (d) and (g) Scatter plots of training sample values (e.g. 500-m pixel) derived from averaging across smaller units (e.g. 30-m pixels) with (Inferred mean) and without (True mean) pre-average binning. (b), (e) and (h) The histogram counts of the sample values. (c), (f) and (i) The difference in histogram counts. The inputs are 1000 simulated Beta distributions made to fit a tree cover range of 0–80% (a, b and c), 0–100% (d, e and f) and 20–100% (g, h and i), respectively, with the Beta shape variables α and β varying between 0.5 and 6

From these, training pixel percentage tree cover values were derived from averaging 20×20 grid cells (e.g. c. 30-m Landsat pixels) into larger sized tiles (e.g. 500-m MODIS pixels). Tree covers were either binned or kept unbinned prior to the averaging into tiles. Again, we reasoned that if there was no bias introduced through binning, the variation in percentage tree cover over the resulting training pixels would be similar without a skew towards specific values of percentage tree cover. We focused on the effects of spatial autocorrelation, which is inherent in most landscapes (Gomez-Sanz, Bunce, & Elena-Rossello, 2014) and in real tree cover gradients, by modelling and comparing (i) fully random uniform and Beta distributions of tree cover, (ii) a sharp boundary landscape, in which half of the landscape was assigned a 100% tree cover and the remaining half a 0% tree cover, (iii) a patchy landscape of autocorrelated Beta distributions across the 0–100% tree

cover range, (iv) a linear gradient of tree cover decreasing from 100% to 0% tree cover along the landscape x axis, and (v) a complex gradient representing Beta distributions across the 0–100% tree cover range. More detail about the Monte Carlo simulations with example landscapes is given in Appendix S2.

Noting, first, that when the two-dimensional simulations were set up to produce the same random non-spatially autocorrelated landscapes as in the semi-analytical experiment, the Monte Carlo approach gives almost identical results as the application of Equation 1 (Fig. S2.4). It further seems that there is the introduction of a second strong bias effect in the case of landscapes with a tree cover gradient (Figure 2). This result is considered in more detail in Appendix S2 and shows that in such a situation pre-average binning causes the training pixel values to inevitably converge towards the bin-class averages. In

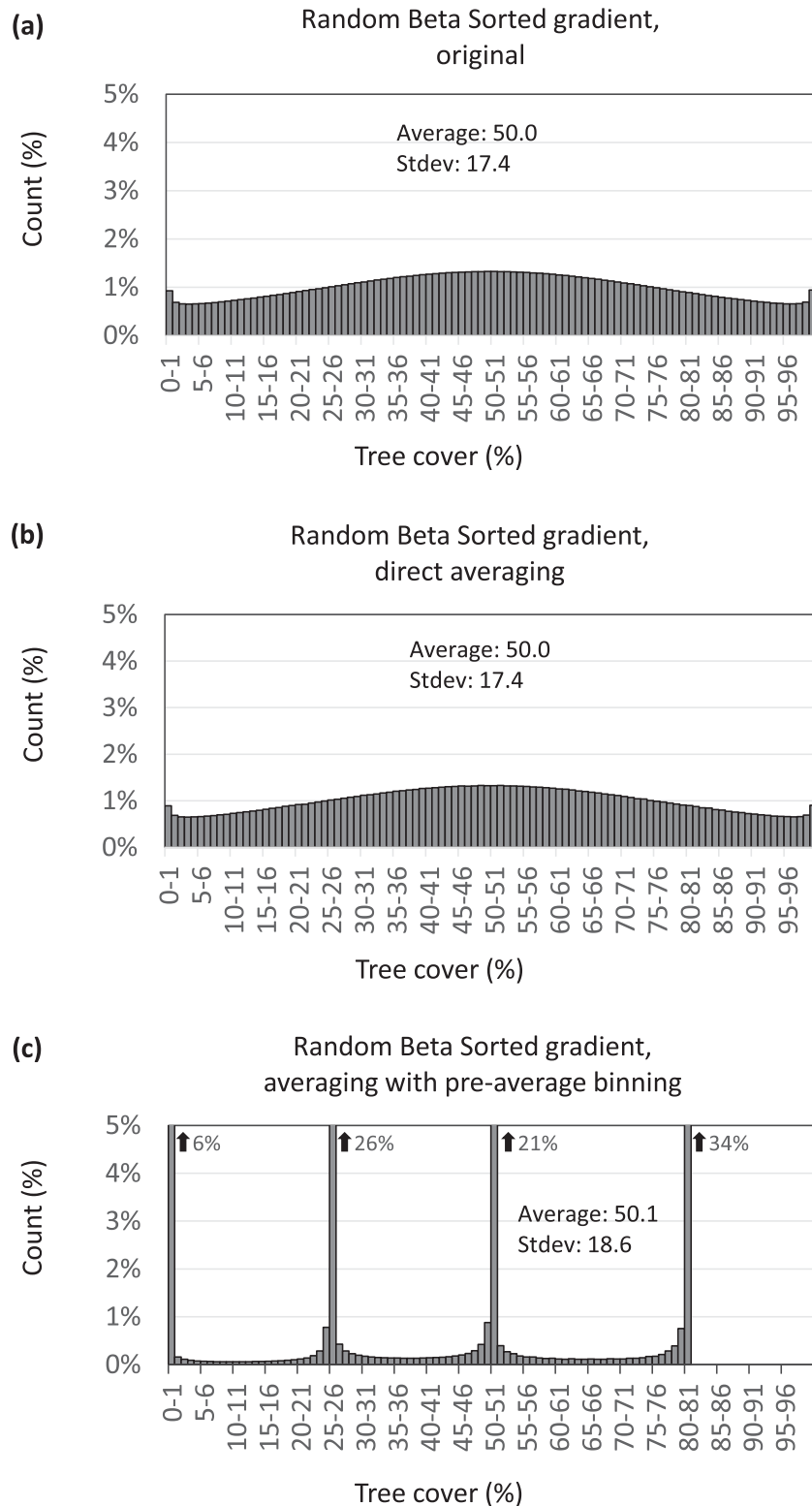


FIGURE 2 (a) Histograms of the original tree cover values and the training sample values (e.g. 500-m pixel) acquired through averaging (b) without pre-average binning or (c) with pre-average binning. The original values are from Monte Carlo simulated theoretical two-dimensional landscapes showing a complex gradient defined using Beta distributions (see Appendix S2 for more detail)

other words, in gradually changing landscapes, the binning procedure considerably lowers the variation in percentage tree cover. This results in a dominant proportion of training pixels with bin-class average values of 0%, 25%, 50% and 80%. Hence, binning tree cover values

before averaging will inevitably underrepresent spatial variation and gradual transitions in tree cover.

Having been trained by biased samples, the CART model must necessarily propagate these biases to the resulting percentage tree

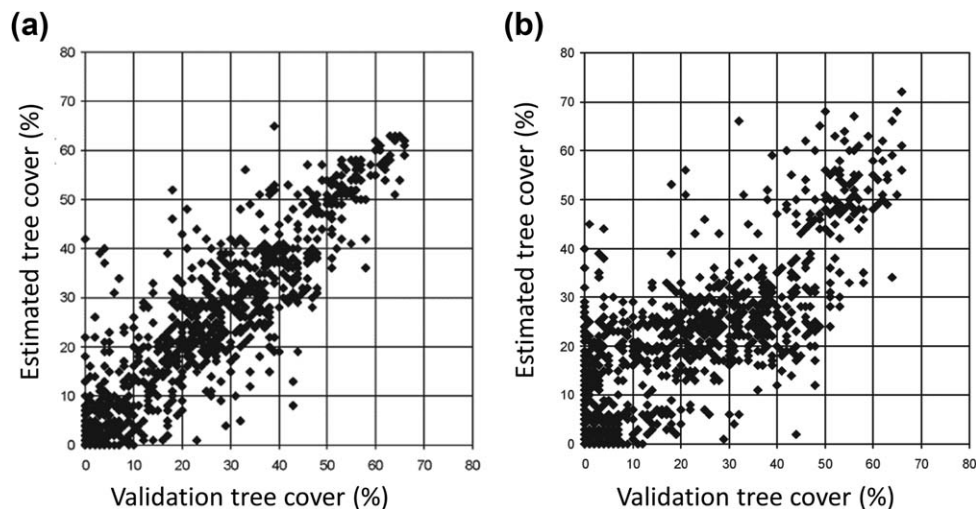


FIGURE 3 (a) Fig. 7(d) of Hansen et al. (2005) showing the validation results for tree cover estimates predicted by one of the four CART variants that were tested for Zambia. All four variants used training sample values representing averages without pre-average binning, which resulted in very similar validation plots, one of which is shown here. (b) Fig. 12 of Hansen et al. (2005) showing the validation results for the tree covers of the MODIS VCF collection 3 product using the same Zambia reference data as in (a). Here a CART was using training sample values representing averages with pre-average binning

cover map. However, because of (i) the complexity of real landscapes, (ii) the large number of nodes in the model, (iii) the introduction of a stepwise regression at each node to deliver a more contiguous range of values (Hansen, DeFries, Townshend, Sohlberg, et al., 2002), and (iv) the variability in the uncertainties of the predicted tree cover, this bias will not always be immediately apparent. Of the published data we know of, the clearest evidence of a bias can be observed in Hansen et al. (2005) and explains the 38–45% discontinuity shown in fig. 1(b) of Staver and Hansen (2015). Hansen et al. (2005) produced, applying the VCF approach, 500-m percentage tree cover maps for a region in Zambia using 500-m training samples that did not undergo pre-average binning at the 30-m Thematic Mapper pixel scale. Comparison with independent validation data show that, in this case, the CART variants being tested were able to reproduce the continuous tree cover gradient observed in the validation data (Figure 3a). But when the same validation data were used to evaluate the MODIS VCF collection 3 created using training samples that had undergone pre-average binning (Figure 3b and also fig. 1(b) in Staver and Hansen (2015)) a clear reduction in values near the 10% and 40% cover range is revealed, indicating a bias. Similarly, the scattergram in Sexton et al. (2013) that compares the MODIS VCF collection 5 with validation data for three sites in America and one in Costa Rica suggests that there may be a slight frequency minimum around the 60% tree cover range, matching the conditions shown in Figure 1g–i. However, without a statistical test for unimodality (Hartigan & Hartigan, 1985) the data in this case are inconclusive.

The bias will not always be readily noticeable. Moreover, the bias will not exist over areas where the corresponding training samples were derived from areas where there was no gradient and/or the tree cover distributions were symmetrical. As an example, the validation VCF scattergrams for the taiga–tundra transition zone, shown in Mon-

tesano et al. (2009) for the MODIS VCF collection 4, reveal no clear clustering of tree covers around the bin-class averages.

We have demonstrated here, using combined semi-analytical and two-dimensional Monte Carlo simulations and data published by Hansen et al. (2005), that tree cover frequency minima and maxima could be caused by a bias in the samples used to train the CART model. This bias will be present in all MODIS VCF collections in areas where the majority of the corresponding training samples were derived from landscapes that have an asymmetric tree cover distribution, contain a tree cover gradient, or a combination of both. This effect is likely to be further enhanced by the CART through the inherent aggregation of predicted values around nodal means. In support of Hanan et al. (2014), we argue that the MODIS VCF tree cover product should not be used to detect discontinuities in tree cover as, although landscapes vary across the globe, a majority will contain local tree cover distributions that are asymmetric and some will contain tree cover gradients.

ACKNOWLEDGMENT

Part of this work was supported through the NERC project TROBIT (NE/D005469/1).

REFERENCES

- DeFries, R. S., Hansen, M. C., Townshend, J. R. G., Janetos, A. C., & Loveland, T. R. (2000). A new global 1-km dataset of percentage tree cover derived from remote sensing. *Global Change Biology*, 6, 247–254.
- DeFries, R. S., Townshend, J. R. G., & Hansen, M. C. (1999). Continuous fields of vegetation characteristics at the global scale at 1-km resolution. *Journal of Geophysical Research-Atmospheres*, 104, 16911–16923.
- DiMiceli, C. M., Carroll, M. L., Sohlberg, R. A., Huang, C., Hansen, M. C., & Townshend, J. R. G. (2011). Annual global automated MODIS

- vegetation continuous fields (MOD44B) at 250 m spatial resolution for data years beginning day 65, 2000-2010. Collection 5 percent tree cover. College Park, MD: University of Maryland.
- Favier, C., Aleman, J., Bremond, L., Dubois, M. A., Freycon, V., & Yanga-kola, J.-M. (2012). Abrupt shifts in African savanna tree cover along a climatic gradient. *Global Ecology and Biogeography*, 21, 787–797.
- Gomez-Sanz, V., Bunce, R. G. H., & Elena-Rossello, R. (2014). Landscape assessment and monitoring. In J. Azevedo, A. H. Perera, & M. A. Pinto (Eds.), *Forest landscapes and global change* (p. 199). New York: Springer Science + Business Media.
- Hanan, N. P., Tredennick, A. T., Prihodko, L., Bucini, G., & Dohn, J. (2014). Analysis of stable states in global savannas: Is the CART pulling the horse?. *Global Ecology and Biogeography*, 23, 259–263.
- Hanan, N. P., Tredennick, A. T., Prihodko, L., Bucini, G., & Dohn, J. (2015). Analysis of stable states in global savannas – a response to Staver and Hansen. *Global Ecology and Biogeography*, 24, 988–989.
- Hansen, M. C., DeFries, R. S., Townshend, J. R. G., Carroll, M., Dimiceli, C., & Sohlberg, R. A. (2003). Global percent tree cover at a spatial resolution of 500 meters: First results of the MODIS vegetation continuous fields algorithm. *Earth Interactions*, 7, 10.
- Hansen, M. C., DeFries, R. S., Townshend, J. R. G., Marufu, L., & Sohlberg, R. (2002). Development of a MODIS tree cover validation data set for Western Province, Zambia. *Remote Sensing of Environment*, 83, 320–335.
- Hansen, M. C., DeFries, R. S., Townshend, J. R. G., Sohlberg, R., Dimiceli, C., & Carroll, M. (2002). Towards an operational MODIS continuous field of percent tree cover algorithm: Examples using AVHRR and MODIS data. *Remote Sensing of Environment*, 83, 303–319.
- Hansen, M. C., Townshend, J. R. G., DeFries, R. S., & Carroll, M. (2005). Estimation of tree cover using MODIS data at global, continental and regional/local scales. *International Journal of Remote Sensing*, 26, 4359–4380.
- Hartigan, J. A., & Hartigan, P. M. (1985). The DIP test of unimodality. *Annals of Statistics*, 13, 70–84.
- Hirota, M., Holmgren, M., Van Nes, E. H., & Scheffer, M. (2011). Global resilience of tropical forest and savanna to critical transitions. *Science*, 334, 232–235.
- Montesano, P. M., Nelson, R., Sun, G., Margolis, H., Kerber, A., & Ranson, K. J. (2009). MODIS tree cover validation for the circumpolar taiga-tundra transition zone. *Remote Sensing of Environment*, 113, 2130–2141.
- Murphy, B. P., & Bowman, D. M. J. S. (2012). What controls the distribution of tropical forest and savanna? *Ecology Letters*, 15, 748–758.
- Scheffer, M., Hirota, M., Holmgren, M., Van Nes, E. H., & Chapin, F. S. (2012). Thresholds for boreal biome transitions. *Proceedings of the National Academy of Sciences USA*, 109, 21384–21389.
- Sexton, J. O., Song, X.-P., Feng, M., Noojipady, P., Anand, A., Huang, C., ... Townshend, J. R. (2013). Global, 30-m resolution continuous fields of tree cover: Landsat-based rescaling of MODIS vegetation continuous fields with lidar-based estimates of error. *International Journal of Digital Earth*, 6, 427–448.
- Staver, A. C., Archibald, S., & Levin, S. A. (2011). The global extent and determinants of savanna and forest as alternative biome states. *Science*, 334, 230–232.
- Staver, A. C., & Hansen, M. C. (2015). Analysis of stable states in global savannas: Is the CART pulling the horse? – a comment. *Global Ecology and Biogeography*, 24, 985–987.
- Taboga, M. (2012). *Lectures on probability theory and mathematical statistics*. CreateSpace Independent Publishing Platform.
- Townshend, J. R. G., Hansen, M. C., Carroll, M. L., DiMiceli, C. M., Huang, C., & Sohlberg, R. A. (2011). *User guide for the MODIS vegetation continuous fields product. Collection 5 version 1* (p. 12). College Park, MD: University of Maryland.
- Xu, C., Hantson, S., Holmgren, M., van Nes, E. H., Staal, A., & Scheffer, M. (2016). Remotely sensed canopy height reveals three pantropical ecosystem states. *Ecology*, 97, 2518–2521.

BIOSKETCH

FRANCE GERARD is a senior scientist at the Centre for Ecology and Hydrology, UK. She has 25 years of experience in Earth observation. Her research covers remote sensing of forest and grasslands and characterizing vegetation dynamics and phenology in a wide range of environments. Her work focuses on the development of Earth observation-derived data to constrain and validate models, support process understanding and monitor vegetation. She is currently working with unmanned aerial vehicle, hyperspectral, camera and radar observations to determine habitat condition and better understand the remotely sensed vegetation phenology signals.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Gerard F, Hooftman D, van Langevelde F, Veenendaal E, White SM, Lloyd J. MODIS VCF should not be used to detect discontinuities in tree cover due to binning bias. A comment on Hanan et al. (2014) and Staver and Hansen (2015). *Global Ecol Biogeogr*. 2017;26:854–859. <https://doi.org/10.1111/geb.12592>