ChatGPT

Sample Selection in Tropical Deforestation Analyses

Limiting Analysis to Forest-Rich Regions

One common approach in tropical deforestation research is to restrict the sample to areas with significant forest cover at baseline. Many studies set a minimum forest cover threshold (often based on Hansen's treecover2000 data with ≥30% canopy density) to define "forested" units for analysis. For example, a study in Uganda limited its sample to grid cells that had at least 30% forest cover in 2000 before evaluating deforestation outcomes [1]. Similarly, research on infrastructure impacts in Cambodia restricted the analysis to cells classified as initially forested (e.g. ≥10% canopy cover in year 2000) in order to focus only on areas where deforestation could plausibly occur [2]. In the Brazilian Amazon, policy evaluation studies have employed even higher cutoffs – one robustness check used only municipalities with over 50% of their territory in native forest as of the pre-policy year [3] [4]. By excluding urbanized or arid districts with little to no natural forest, these studies ensure the outcome (forest loss rate) is meaningful and not dominated by structurally zero values.

Restricting the sample to forest-rich regions has clear practical and statistical justifications. **First**, it improves the relevance of the deforestation rate metric. In districts with negligible forest, a "deforestation rate" is either ill-defined or always zero, which can introduce noise or instability (e.g. a tiny forest patch lost yields an extreme percentage loss). Focusing on units above a forest cover threshold mitigates this problem. **Second**, it addresses heterogeneity in deforestation dynamics. Areas with very low remaining forest often exhibit different dynamics (e.g. a slowing of forest loss due to scarcity) compared to frontier areas with abundant forest [3]. By analyzing only well-forested ADM2 units, researchers can hold the context more constant. For instance, Assunção et al. (2013) note that municipalities with little forest left may respond differently to policies than those still heavily forested, so they tested policy impacts on the high-forest subset to ensure comparability [3]. **Third**, it reduces measurement error and denominator issues in computing deforestation rates. Many studies use Hansen et al.'s treecover data (≥30% canopy cover) as the base forest area; excluding sites below a certain forest percentage avoids ratios based on very small denominators that could bias regression results.

However, **the validity and trade-offs** of this approach must be considered. By excluding low-forest areas, one is essentially selecting on a **pre-treatment characteristic**, which is less problematic than conditioning on outcomes (discussed next) but still affects external validity. The findings will generalize only to forested regions rather than the entire country or all ADM2 units. In Indonesia's case, limiting analysis to districts with (say) >30% natural forest cover in 2019 would drop many urbanized or agricultural districts – the results would pertain to "forested districts" rather than Indonesia as a whole. If those excluded districts are fundamentally different (e.g. higher development, different land-use patterns), the **external validity** of causal estimates is narrower [3]. There is also a mild risk of **selection bias** if baseline forest cover correlates with unobserved factors that also affect deforestation or the treatment of interest. For instance, districts with very low forest might have intensive agriculture and infrastructure, which could be related to both policy implementation and deforestation pressure. Most studies mitigate this by controlling for observable differences or using fixed effects, but it's noted that an arbitrary threshold could omit areas that, while currently low in forest, might have experienced deforestation in the past (a form of survivor bias in the sample). On balance, this method is widely used in tropical forestry research as a reasonable compromise – it improves the **precision and interpretability** of deforestation regressions [2] [3], but researchers make clear that results apply to the forest-rich sub-population.

Excluding Regions with No Deforestation Events

Another approach is to drop observation units that experienced zero deforestation (e.g. no GFW alerts) throughout the study period. The motivation is often to avoid a flood of zero outcomes that could skew model fit or to focus on areas “active” in deforestation. This method is more controversial, as it entails conditioning on the outcome. In practice, some analyses have intentionally or by necessity used this exclusion. For example, a recent study of economic growth in the Amazon **“trimmed”** its sample to municipalities that did experience deforestation, eliminating those with none, in order to examine how varying deforestation levels impacted GDP [5]. By restricting to the 610 municipalities with non-zero deforestation, the authors treated deforestation as a treatment uniformly received in that subsample [6]. Another case comes from an evaluation of Brazil’s priority “blacklist” policy: the dataset excluded municipalities with no deforestation **variation** from 2002–2011 because a normalized deforestation rate could not be computed for those with constant zero deforestation [7]. In other words, if a municipality never had any forest loss in the panel, it was dropped since the statistical model (which centered deforestation by each municipality’s mean) required some change over time [7]. These examples illustrate that, in practice, researchers sometimes remove zero-deforestation units due to modeling constraints or to concentrate on the deforestation frontier.

Despite such usage, **literature generally cautions against outright exclusion of zero-deforestation observations** when the goal is causal inference on drivers of deforestation. Dropping all ADM2 units with no alerts in 2019–2025 means **conditioning on the outcome variable**, which can introduce **selection bias** and distort causal estimates. If areas with zero alerts are systematically different – for instance, predominantly urban districts, or conversely well-protected forest districts – their exclusion biases the sample toward high-pressure locales. Any factor that helped keep deforestation at zero (e.g. strong protection policies or remoteness) would be omitted from analysis, potentially overstating the effect of other drivers in the remaining sample. In a PLOS ONE study on Amazon governance, reviewers explicitly noted that the authors should include municipalities with zero deforestation rather than exclude them, to avoid giving the impression of cherry-picking favorable cases [8]. The authors clarified that they indeed included all municipalities with available deforestation data, ensuring that zeros were treated as valid outcomes [9]. This peer-review comment highlights the normative view: excluding zero-deforestation areas without very strong justification is frowned upon, as it threatens the **integrity of causal analysis**.

From a statistical standpoint, removing all zeros can lead to **selection on unobservables**. Imagine an instrumental variable (IV) analysis where the instrument’s effect is partly to reduce deforestation to zero in some districts. If those districts are dropped, the IV estimation would ignore instances where the instrument successfully prevented deforestation, thereby biasing the estimated treatment effect. Similarly, OLS or panel regressions would be estimated only on the subset of units that were prone to deforest, possibly yielding different coefficient magnitudes than one would get with the full sample. In short, the sample becomes conditional on deforestation having occurred – which can alter relationships. For example, one study found that focusing only on deforesting municipalities changed the interpretation to a “treated-only” effect, and they noted no control group remained after trimming [6]. Unless the research question specifically targets the conditional process of how much forest is lost given that some loss occurs, it is usually more appropriate to model the many zero outcomes rather than drop them. Techniques like Tobit or zero-inflated models, or two-part models (first modeling the probability of any deforestation, then the amount conditional on occurrence), are alternatives discussed in econometric literature when zeros are frequent. These approaches retain the full sample and thus preserve **external validity** – allowing one to say something about all regions, including those that stayed intact.

In tropical countries including Indonesia, many districts might report no forest loss in a short window simply because they have stable forest or very little forest to begin with. Excluding such cases outright risks overlooking important “control” observations where deforestation pressure was absent or successfully mitigated. Indeed, by analyzing why some areas had zero loss (perhaps due to policies or community stewardship), researchers gain insight into causal mechanisms. That said, there are scenarios where excluding zeros is defensible: for instance, when outcome data are unreliable below a certain threshold or when focusing on the intensive margin of deforestation (conditional on some being observed). The key is to acknowledge the **trade-off**. Dropping no-deforestation ADM2 units sharpens the focus on active deforestation dynamics but at the possible cost of **selection bias** (conditioning on a post-treatment outcome) and reduced generalizability beyond the high-deforestation sub-sample [8] .

Bias and Validity Considerations in Causal Inference

Both methods of sample restriction have implications for causal inference studies of deforestation. The goal in an IV or regression analysis is typically to obtain an unbiased estimate of a treatment (or driver) effect on deforestation, and to ensure the findings apply to a relevant population. Below we summarize how each approach influences bias and validity:

- **Statistical Bias:** Limiting the sample to forested areas (Approach 1) is **less likely to introduce endogenous bias** since baseline forest cover is predetermined. It may, however, omit some heterogeneity that could confound results if not properly controlled (e.g. forest cover correlating with socio-economic factors). Most studies handle this by including covariates or fixed effects that account for differences in baseline forest extent [3] . Excluding zero-deforestation units (Approach 2) raises a bigger red flag: it constitutes **selection on the dependent variable**, which can bias coefficients. Essentially, the regression no longer represents factors affecting the probability of deforestation versus none, but only the intensity where deforestation happened. Unless the analytic focus is explicitly conditional, this can violate the assumptions of OLS/IV by conditioning on a collider (the occurrence of deforestation). The Brazil blacklist study’s technical paper acknowledged this issue by normalizing deforestation and dropping constant-zero cases strictly for computational reasons, not because zeros were unimportant [7] .

- **External Validity:** Approach 1 sacrifices some breadth of applicability in exchange for precision. The causal estimates derived from a high-forest subset may not extrapolate to low-forest regions. For example, an IV estimate of the effect of protected areas on deforestation, if run only on dense-forest districts, tells us about those districts but not about sparsely forested ones (which might respond differently or not at all to protection). Researchers are aware of this and sometimes present it as a scope condition (e.g. “our results apply to the forest zone of Indonesia”). In contrast, Approach 2 dramatically narrows external validity: conclusions apply only to places that actually experienced deforestation. In an extreme sense, it answers a different question – “what drives variation in deforestation among deforesting areas?” rather than “what drives deforestation across all areas?” If policymakers need country-wide insight (including why some places stayed forested), a zero-only sample is incomplete. This concern is reflected in the peer review feedback urging inclusion of all municipalities in an Amazon analysis to maintain comprehensive insight [8] .

- **Selection Bias and Causal Consistency:** In causal inference, one worries that the sample selection correlates with treatment or potential outcomes. With a forest-cover filter (Approach 1), the selection is based on a relatively exogenous trait; however, if the treatment of interest (say, an anti-logging policy or an economic shock) disproportionately occurs in excluded vs. included areas, one must check for any resulting bias. Often, studies will report robustness checks: e.g. confirming that results hold when including more observations or interacting the treatment with a baseline forest measure [4]. With outcome-based exclusion (Approach 2), selection bias is more acute. For instance, if higher governance quality leads both to fewer deforestation alerts (hence more zeros) and to lower deforestation rates where it does occur, excluding the zero cases associated with good governance skews the estimated effect of governance on deforestation. In IV terms, dropping zeros can break the monotonicity or exclude "always-takers"/"never-takers" groups, affecting the LATE (Local Average Treatment Effect) interpretation. Researchers who have used Approach 2 are typically careful to note this. In the Amazon GDP study, the authors treated their trimmed sample analysis as a supplementary insight (highlighting non-linear effects in deforesting municipalities) and still presented full-sample analyses elsewhere [10] [6]. This underscores that one should not casually drop zero-outcome observations in causal research without addressing how it might bias results.

In summary, **precedent from tropical forest literature suggests**: (1) Applying a baseline forest cover criterion is a widely accepted practice to improve the focus and reliability of deforestation analyses, so long as one acknowledges the reduced generality of the findings [3] [4]. It can actually enhance internal validity by making treated and control units more comparable in terms of having forest at risk. (2) Excluding areas with no deforestation events is generally viewed with caution. While there are cases where it is done (due to data or method constraints), researchers warn of potential biases and often prefer to incorporate zero-deforestation observations either by modeling them separately or by using appropriate statistical techniques [8] [7]. In the context of Indonesia or other tropical regions (Amazon, Congo Basin, etc.), most causal studies retain zero-loss areas in the sample and use them to help identify effects (for example, showing a policy kept deforestation at zero in some places). Only when focusing on a very specific conditional relationship or when zero inflation makes standard regression untenable do they resort to trimming the sample, and even then with robustness checks. The balance of evidence from prior research is that **sample selection should align with the research question**: if one is evaluating drivers of deforestation, including all forested and non-deforested units improves credibility, whereas if one is examining impacts of deforestation (as a treatment), a conditional sample might be justified but must be interpreted carefully in light of potential selection bias.

Sources: Tropical deforestation studies and reviews, including methods from Indonesia, Amazonia, and global analyses [3] [4] [6] [8] [7], as detailed above.

---

[1] [PDF] on the ground - cifor-icraf
https://www.cifor-icraf.org/publications/pdf_files/Books/BCIFOR1403.pdf

[2] [PDF] Environmental Impacts of Chinese Government-Funded ... - AidData
https://docs.aiddata.org/ad4/pdfs/WPS114_Environmental_Impacts_of_Chinese_Government-Funded_Infrastructure_Projects__Evidence_from_Road_Building_in_Cambodia.pdf

[3] climatepolicyinitiative.org
https://climatepolicyinitiative.org/wp-content/uploads/2013/05/DETERring-Deforestation-in-the-Brazilian-Amazon-Environmental-Monitoring-and-Law-Enforcement-Technical-Paper.pdf

[4] [7] climatepolicyinitiative.org
https://climatepolicyinitiative.org/wp-content/uploads/2014/08/Getting-Greener-by-Going-Black-Technical-Paper.pdf

[5] [6] [10] Deforestation and economic growth in the Amazon region: investigating with a transposed environmental Kuznets curve - Ecology & Society
https://ecologyandsociety.org/vol30/iss4/art29/

[8] [9] What’s governance got to do with it? Examining the relationship between governance and deforestation in the Brazilian Amazon | PLOS ONE
https://journals.plos.org/plosone/article/peerReview?id=10.1371/journal.pone.0269729