

Supplementary Information: Putting numbers on timber illegality risk: the case of *ipê* in Pará

Caroline S. S. França, U. Martin Persson, Tomás Carvalho
and Marco Lentini

	001
	002
	003
	004
	005
	006
	007
	008
	009
	010
	011
	012
	013
	014
	015
	016
	017
	018
	019
Supplementary Figures	2
Figure S1 Temporal evolution of production by municipality	2
Figure S2 Contrasting species yield between logging permits and RADAM national inventory	3
Figure S3 Origin of overstated yields	3
Figure S4 Yield, Volume and Area distribution by status of logging permit	4
Figure S5 Overview of flows between systems	5
	020
	021
	022
	023
	024
	025
	026
	027
	028
Supplementary Text	6
Combining timber origin control systems	6
	029
	030
	031
	032
	033
	034
	035
	036
	037
	038
	039
	040
	041
	042
	043
	044
	045
	046

Supplementary Figures

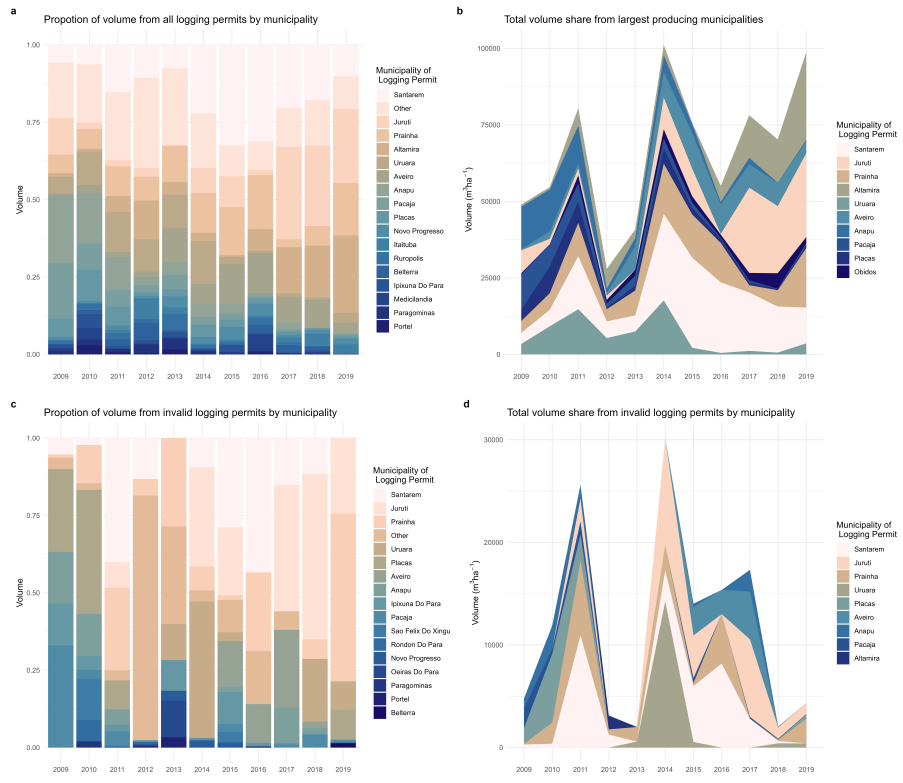


Fig. S1 Temporal evolution of production by municipality for volume originating from all logging permits (a-b) versus only for invalid ones (c-d).

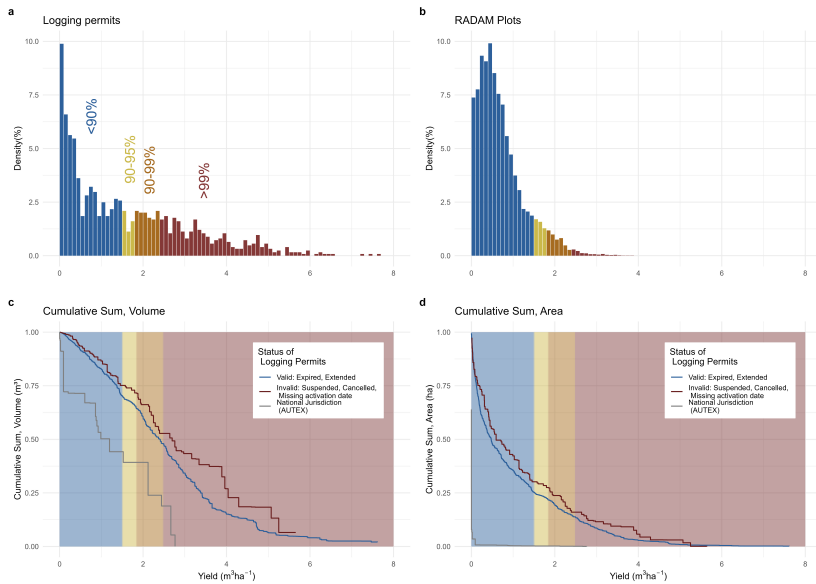


Fig. S2 Contrasting species yield between volume entering the supply chain and RADAM national inventory. *Ipê* yield density for our data (a) in contrast to the national RADAM forest inventory (b). Colors show the 90th, 95th and 99th percentiles in the RADAM inventory (all panels). Panels (c) and (d) show the difference in cumulative distribution by volume and area, respectively, for logging permits of different status, highlighting that the large share of total volume (48%) that is in the 99th percentile yield comes from a relatively small area (14%).

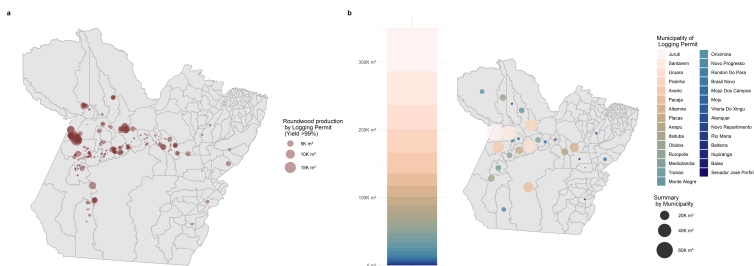


Fig. S3 Origin of overstated yields. Location of logging permits with yield above the 99th percentile class (a). Volume from overstated permits entering the supply chain summarized by municipality of origin (b).

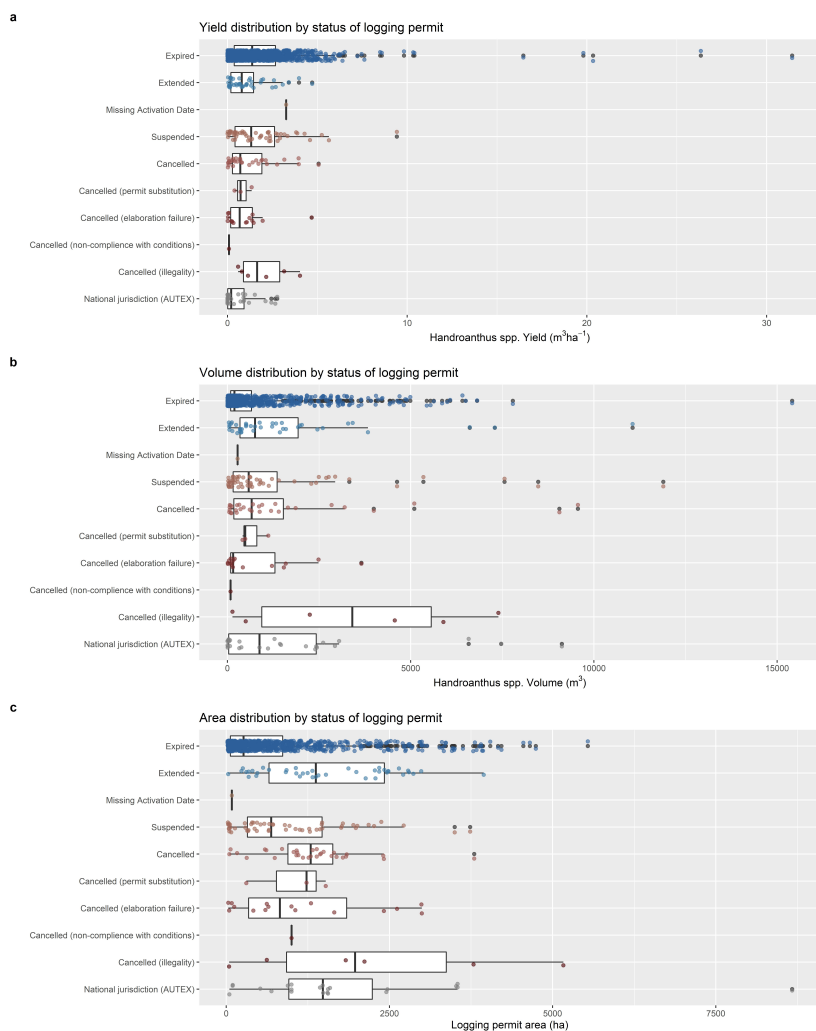


Fig. S4 Yield (a), volume (b) and area (c) distributions by status of logging permits. Area distributions in (c) does not display seven logging permits with areas above 10,000 ha, all under the "National Jurisdiction (AUTEX)" category.

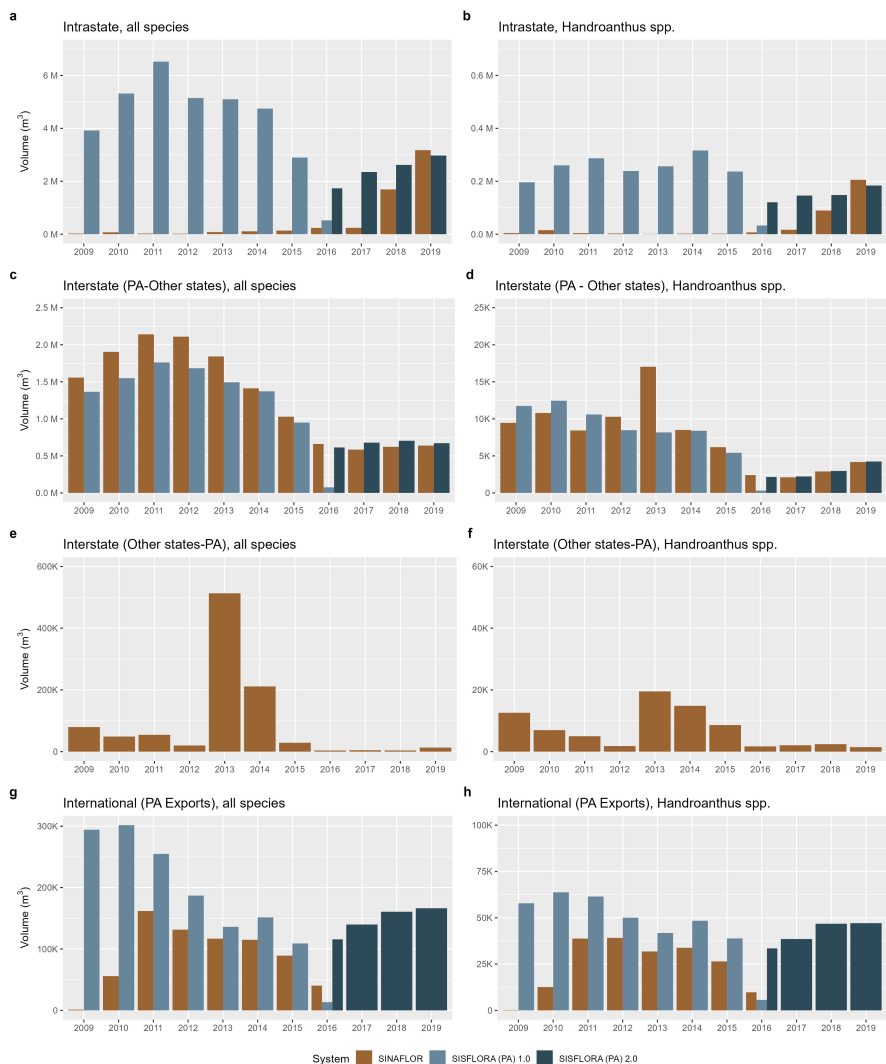


Fig. S5 Summary of total flows in the three different reporting systems: SINAFLO (national-level), SISFLORA (Pará state-level system) 1.0 (pre-2016) and 2.0 (post-2016). Panels a, c, f, g show an overview of transacted timber products for all species, and panels b, d, f, h for *Handroanthus* spp. (and *Tabebuia* spp. now *Handroanthus*, see methods) alone, at distinct jurisdictional levels. No conversion has been applied to at this stage.

Supplementary Text

Combining timber origin control systems

Constructing a comprehensive and consistent dataset on timber transactions of the state of Pará requires cleaning and harmonizing data from three different datasets: the state-level (SISFLORA-PA) system prior to and after 2016 (versions 1.0 and 2.0, respectively), as well as the national-level (SINAFLOR) system. The upgrade of the SISFLORA-PA system in 2016 primarily allowed for the inclusion of more detailed stump and log tracing via RFID-based chip and bar code system, but also other system refinements. This has resulted in two slightly differing data structures, with the inclusion of new variables (e.g. log bar code, type of destination) on the one hand, and suppression (e.g. type of origin) on the other. Such differences are, however, minor when comparing to those between SISFLORA-PA and SINAFLOR, which includes not only differences in data structure, but also implications related to the jurisdiction where norms and procedures apply. Thus aside from data structure, the overlapping policyscape translates into reporting duplication between systems which has so far not been properly addressed by SINAFLOR.

In attempting to remove duplicate data entries (i.e., transport records that have been reported in both the SINAFLOR and SISFLORA systems), we found no optimal solution based on a set of unique identifiers drawn from the data collected and made available to us. Aside from the lack of a unique global key [1]—which could plausibly be expected from SINAFLOR given their institutional mandate of integration between systems—data on CPFs for actors of origin and destination have been partially anonymized in the SINAFLOR data, in line with recent national data protection law (LPGD) [2]. It is worth clarifying these differ from CNPJ, which are made openly available (e.g.

for a data sample: [https://dadosabertos.ibama.gov.br/dados/DOF/PA/trans-
porte/2021.html](https://dadosabertos.ibama.gov.br/dados/DOF/PA/trans-
porte/2021.html)). While there are benefits to individual’s data protection
(which can be done with methods that provide individual anonymization while
still maintaining information of timber origin and commercialization), given
the remaining variables that are present in both datasets, the CPF is important
in supporting identification in the uniqueness of a given transaction through-
out the period of analysis. This is particularly true moving from intrastate-only
to national-level control: where there are needs for alignments of state-
level cadasters with environmental authorities (e.g. CEPROF, *Cadastro de
Exploradores e Consumidores de Produtos Florestais* or ”Cadaster of Forest
Products Exploiters and Consumers”) to those of national-level cadasters (e.g.
CTF, *Cadastro Técnico Federal* or ”Federal Technical Cadaster”). In this
study, we used the partial anonymization implemented in the SINAFLOR-
level data (i.e. making available only the first three digits of CPF and only the
initial of names associated with CPFs) in order to establish actor’s unique-
ness between SINAFLOR and SISFLORA-PA systems. First, while partially
anonymized in SINAFLOR, details added to other variables (e.g. yard of ori-
gin and destination) still in SINAFLOR could be used to determine full CPF
and individual identity (here a weakness of the anonymization method used).
Last, by combining the first three digits of CPF and initials, actor uniqueness
can still be validated with SISFLORA-PA data. Moreover, while we make no
sensitive data available, individual’s cadaster (i.e. CPFs) are an important key
for deduplication given the lack of proper data harmonization from the part
of government authorities, and more specifically SINAFLOR.

Aside from the lack of global unique keys for removing duplicates between
systems, further challenges can be better illustrated when systematic group-
ing multiple variables. Firstly, only testing the amount of roundwood for *ipê*

(*Handroanthus* spp.) transported in 2019 (a year that broadly appears to have higher level of equal reporting between systems and thus higher duplication level, see Fig. S6b for example), SINAFLOR reports a total of 138,455 m³ of which 79% state an origin that is equal to destination. When looking at the same parameters for SISFLORA-PA, volume totals 114,671 m³ (a lower value can be expected, as SINAFLOR may encompass reporting for licensed enterprises at both state- and national-level), but with only 7% of actor of origin being the same as destination. Such large differences in reporting key information as origin and destination suggests intrinsic differences between datasets. Indeed, most SINAFLOR records for this year displayed four key features in common: they (i) did not have a listed permit (needed for transport from forest of origin), (ii) had actor of origin that equaled that of destination, (iii) had data on type of origin missing, and (iv) did not display coordinates of origin or destination. Thus, it is likely the flows that displayed this shared lack of reporting are the duplicates from state-level system. Given flows with origin that equal destination have no implication on the physical balance of the input-output analysis (see methods as well), we removed these, which by default minimize the potential double-counting. After the removal of records which displayed common features of duplicates (i.e. systematic lack of many data fields in SINAFLOR) we ensured no clear duplicates between systems remained. We summed up volumes per actor, municipality, product, specie and year and removed from SINAFLOR all instances where the sums were equal between the two systems.

References

- [1] Elmagarmid, A. K. *et al.* Duplicate Record Detection: A Survey. *IEEE Transactions on Knowledge and Data Engineering* **19** (1), 1–16 (2007).

https://doi.org/10.1109/TKDE.2007.250581 .	369
	370
[2] Brazil. Law N ^o 13.709, 14.08.2018 (2016). Available at: https://www.	371
	372
planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/113709.htm .	373
	374
	375
	376
	377
	378
	379
	380
	381
	382
	383
	384
	385
	386
	387
	388
	389
	390
	391
	392
	393
	394
	395
	396
	397
	398
	399
	400
	401
	402
	403
	404
	405
	406
	407
	408
	409
	410
	411
	412
	413
	414