# Beyond the Mean: A Distributional Intensity Metric for Model-Checking Coverage

Follow-up study #5 to Kumazawa et al., PAAMS 2023

2026


**Abstract**

[1] proposes two coverage metrics: diversity $AC_{div}$ (fraction of abstract states visited) and intensity $AC_{int}$ (visits per abstract state), reported as an average over abstract states. We show that the *averaged* $AC_{int}$ is, by construction, the visited-state count divided by a constant, so it is informationally redundant with the visited-state curve and — worse — it *rewards redundant re-visiting.* We propose measuring intensity as the *distribution* of per-abstract visit counts and summarising it with the Gini coefficient and normalised entropy. On reproduced data the distributional metric overturns the average's verdict: a method that visits 2.3× *fewer* states nonetheless covers *more* abstract states and more evenly (Gini 0.556 vs 0.576), exactly the behaviour a coverage practitioner wants. The metric is a drop-in, more discriminating replacement. Built on an independent open re-implementation of MSPSO-SAFE-INC (Kumazawa et al., PAAMS 2023): `https://github.com/yaskodama/kuma2023-repro`.


## 1 The problem with averaged intensity

[1] defines $AC_{int}(s)$ as the number of visited product states whose abstract state is $s$, and plots its average over abstract states. But

$$\frac{1}{|S_{sp}|}\sum_{s} AC_{int}(s) = \frac{\#\text{visited product states}}{|S_{sp}|},$$

i.e. the averaged intensity is the visited-state count up to a constant factor. On our reproduced runs the Pearson correlation between averaged $AC_{int}$ and the visited-state count is 1.000 — the "intensity" panel carries no information the "visited states" panel does not. Two further problems follow: the average (i) cannot tell whether visits are spread evenly or piled onto a few abstract states, and (ii) *increases when search wastefully re-visits already covered states*, the opposite of good coverage behaviour.

## 2 A distributional intensity metric

Let $c(s)$ be the visit count of abstract state $s$. We treat $\{c(s)\}_{s\in S_{sp}}$ as a distribution and report two scale-free summaries:

$$\text{Gini} = \frac{\sum_i \sum_j |c_i - c_j|}{2\,|S_{sp}|\sum_i c_i}, \qquad H = \frac{-\sum_s p(s)\log p(s)}{\log |S_{sp}|},\ \ p(s) = \frac{c(s)}{\sum c}.$$

Gini $= 0$ and $H = 1$ mean perfectly even intensity (the ideal: breadth without redundant pile-up); high Gini / low $H$ mean a few abstract states absorb most visits. The Lorenz curve (Fig. 1) visualises the same concentration.

## 3 Setup

Reproduced benchmark, $n$=1000, $p$=5, faithful Table-1 budget, 120 outer iterations. We compare the paper's restart (*none*) with the best method from follow-ups #1/#3 (*CGU+persistent*).

## 4 Results

Table 1 is the crux. CGU+persistent visits 2.3× *fewer* product states than baseline (66,040 vs 154,778), so its averaged $AC_{int}$ is far lower (66.0 vs 154.6) — the paper's metric would call it the *worse* explorer. Yet it covers *more* abstract states (885 vs 881) and distributes visits more evenly (Gini 0.556 vs 0.576, entropy 0.909 vs 0.903). The distributional metric correctly identifies it as the better, less redundant explorer; the average inverts the judgement.

**Table 1:** Intensity measured as average vs as a distribution ($p$=5).

| | none (paper) | CGU+persistent |
|---|---|---|
| visited product states | 154,778 | 66,040 |
| abstract states covered | 881 | **885** |
| averaged $AC_{int}$ (paper) | 154.6 | 66.0 |
| Gini (lower = more even) | 0.576 | **0.556** |
| norm. entropy (higher = more even) | 0.903 | **0.909** |

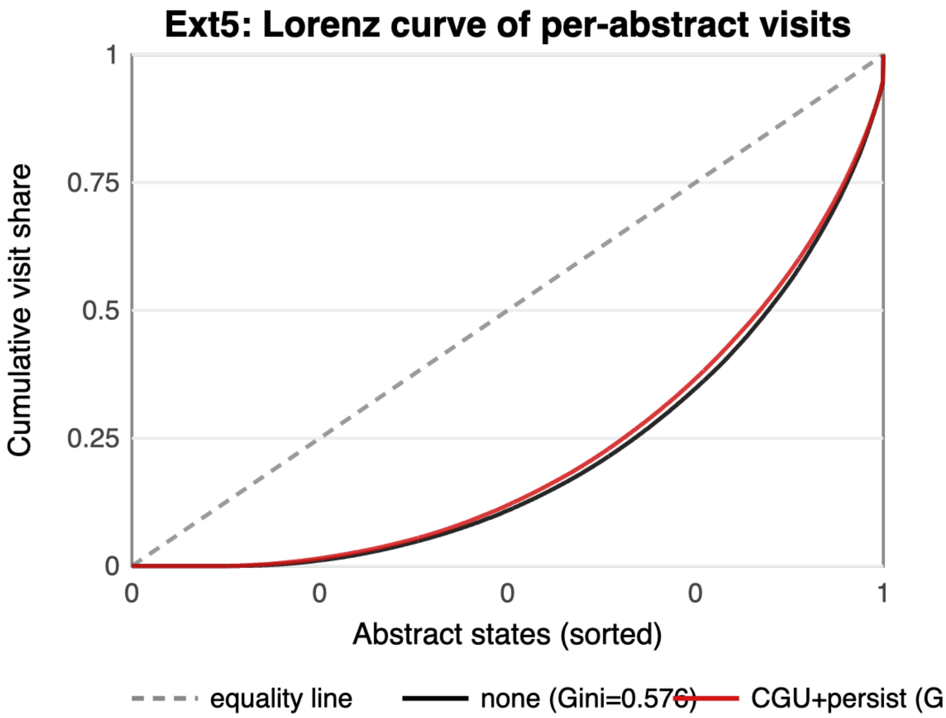


**Figure 1:** Lorenz curves of per-abstract visit counts ($p$=5). A curve closer to the diagonal means visits are spread more evenly across abstract states.

## 5 Discussion

Reporting intensity as a distribution restores the information the average throws away and aligns the metric with the goal: a good model checker covers many abstract states *without* wastefully re-deepening a few. Gini/entropy are cheap (computed from the same visit counts the paper already tracks), scale-free (comparable across instances of different size), and — unlike the average — not monotone in raw effort. The Lorenz curve additionally localises which abstract regions are over- or under-intensified, a diagnostic for steering search.

## 6 Conclusion

The paper's averaged $AC_{int}$ is redundant with the visited-state count and can reward wasteful re-visiting. A distributional intensity metric (Gini, normalised entropy, Lorenz curve) is a drop-in, more discriminating alternative that, on our data, reverses the average's verdict in favour of the genuinely better explorer.

## References

[1] T. Kumazawa, M. Takimoto, Y. Kodama, Y. Kambayashi. *Enhancing Safety Checking Coverage with Multi-swarm Particle Swarm Optimization.* PAAMS 2023.

[2] R. Castaño et al. *Verification Coverage.* arXiv:1706.03796, 2017.