Supplementary Information for "Separating local scale dependence from regional mixing in urban block-size distributions"

Hiroyuki Shima[1*] and Yuri Akiba[2]

[1*]Department of Environmental Sciences, University of Yamanashi, 4-4-37 Takeda, Kofu, 400-8510, Japan.
[2]Mount Fuji Research Institute, Yamanashi Prefectural Government, 5597-1 Kenmarubi, Kamiyoshida, Fujiyoshida, 403-0005, Japan.

*Corresponding author(s). E-mail(s): hshima@yamanashi.ac.jp;
Contributing authors: akiyuri@mfri.pref.yamanashi.jp;

S1 Regularity and exact identities

Let $A = A_* e^{\ell}$. It is sufficient that, for almost every component r, $f(A_* e^{\ell} \mid r)$ is twice differentiable on each compact ℓ-interval and that f, $\partial_\ell f$, and $\partial_\ell^2 f$ possess integrable envelopes. Regional observables and their derivatives must be integrable under the size-conditioned measure. For an area-dependent observable $u(r, A)$, sufficient additional envelopes are $|u|f$, $|u\partial_\ell f|$, and $|\partial_\ell u|f$.

Write $F_r(\ell) = f(A_* e^{\ell} \mid r)$. From $\partial_\ell F_r = -a_r F_r$ and $\partial_\ell f_{\mathrm{agg}} = -a_{\mathrm{agg}} f_{\mathrm{agg}}$, the size-conditioned measure obeys

$$\partial_\ell \rho_A(\mathrm{d}r) = \rho_A(\mathrm{d}r)[-a_r(A) + a_{\mathrm{agg}}(A)]. \tag{S1}$$

Consequently,

$$\begin{aligned}\partial_\ell \mathbb{E}_{\rho_A}[u] &= \mathbb{E}_{\rho_A}[\partial_\ell u] - \mathbb{E}_{\rho_A}[u a_r] + \mathbb{E}_{\rho_A}[u]\mathbb{E}_{\rho_A}[a_r] \\ &= \mathbb{E}_{\rho_A}[\partial_\ell u] - \mathrm{Cov}_{\rho_A}(u, a_r).\end{aligned} \tag{S2}$$

Choosing $u = a_r$ gives the curvature budget; choosing an area-independent u gives the sorting identity.

For nested groups $r = (g, z)$, factor the size-conditioned measure as $\rho_A(\mathrm{d}g, \mathrm{d}z) = \rho_A^G(\mathrm{d}g)\rho_A(\mathrm{d}z \mid g)$. The law of total variance yields

$$\begin{aligned} \partial_\ell a_{\mathrm{agg}} = \mathbb{E}_{\rho_A}[\partial_\ell a_{g,z}] - \mathbb{E}_{\rho_A^G}\left[\mathrm{Var}_{\rho_A(\cdot|g)}(a_{g,z})\right] \\ - \mathrm{Var}_{\rho_A^G}\left(\mathbb{E}_{\rho_A(\cdot|g)}[a_{g,z}]\right). \end{aligned} \tag{S3}$$

The second and third terms are the within-group and between-group parts of the heterogeneity penalty. Regrouping therefore redistributes the penalty without changing pooled curvature.

S2 Continuous pure-power mixtures

Consider

$$f_{\mathrm{agg}}(A) = \int g(\lambda)(A/A_*)^{-\alpha(\lambda)}\mathrm{d}\lambda. \tag{S4}$$

Suppose α has an isolated active minimum α_* at λ_*, and locally

$$\alpha(\lambda) = \alpha_* + c|\lambda - \lambda_*|^m + o(|\lambda - \lambda_*|^m), \tag{S5}$$

$$g(\lambda) = g_*|\lambda - \lambda_*|^\nu[1 + o(1)], \tag{S6}$$

where $c, m, g_* > 0$ and $\nu > -1$. With $L = \ln(A/A_*)$, Laplace's method gives [1]

$$f_{\mathrm{agg}}(A) \sim K(A/A_*)^{-\alpha_*}L^{-\beta}, \qquad \beta = (\nu + 1)/m > 0, \tag{S7}$$

$$a_{\mathrm{agg}}(A) = \alpha_* + \frac{\beta}{\ln(A/A_*)} + o((\ln A)^{-1}). \tag{S8}$$

The aggregate approaches the smallest active exponent from above. Continuous mixing of exact powers therefore also produces asymptotic flattening.

S3 Estimation and numerical audits

The common threshold was $1850\,\mathrm{m}^2$, leaving 51,055 blocks. The main reconstruction used a 260-point geometric grid from 5000 to $50000\,\mathrm{m}^2$, a reflected Gaussian kernel density estimator in log area, and bandwidth 0.35. Bandwidths 0.30 and 0.40 were sensitivity values. Analytic first and second derivatives gave local exponents and curvatures. Bootstrap deviations used log-area bins of width 0.004 and were recentered on the unbinned estimates. Five hundred stratified replicates resampled blocks within wards while preserving ward counts.

Independent finite differences on synthetic pure-power and curved components gave all identity errors below 2×10^{-10}. The maximum nested coarse-graining mismatch over two to four RDN strata and three bandwidths was below 10^{-15}. For the Tokyo sorting identity, the main bandwidth gave maximum absolute and root-mean-square discrepancies 7.33×10^{-5} and 2.93×10^{-5}. Across all 69 bandwidth–leave-one-ward-out cases, the maximum absolute discrepancy was 1.36×10^{-4}.

S4 Synthetic recovery and ward-cluster uncertainty

The synthetic recovery test retained the observed 23 ward sample sizes, the common threshold $A_* = 1850\,\mathrm{m}^2$, the 260-point grid, and bandwidth 0.35. In log area $y = \ln(A/A_*)$, the exact-power null used exponential component densities with rates equal to the 23 fitted Pareto exponents minus one. The positive control used

$$q_i(y) \propto \exp[-r_i y - \eta(e^y - 1)], \qquad \eta = 0.06, \tag{S9}$$

so its known regional local curvature is $\eta e^y = \eta A/A_*$. The value of η was fixed before simulation and was not fitted to the Tokyo curvature. Five hundred datasets per scenario were generated with fixed ward counts and reconstructed with the same reflected-kernel procedure as the empirical data.

Near $10^4\,\mathrm{m}^2$, the exact-power null has analytic pooled curvature -0.0237. Its reconstructed median was -0.0253, with a 95% simulation interval $[-0.155, 0.116]$. Because this null curvature is close to zero, 35.2% of unthresholded point estimates were positive; the sign of a single noisy estimate is therefore not itself a calibrated test. The empirical value 0.246 nevertheless exceeded all 500 null estimates, giving the finite-simulation one-sided result $p = (0 + 1)/(500 + 1) = 0.0020$. For the tempered positive control, the analytic value was 0.300, the reconstructed median was 0.255 [0.107, 0.416], and all 500 estimates were positive. Median curve RMSE over 7000–30000 m^2 was 0.0885 for the exact-power null and 0.147 for the positive control. Decomposition closure errors were below 9×10^{-15}.

Spatial coordinates and block adjacency are not present in the analysis table. We therefore added a whole-ward cluster bootstrap rather than claiming a distance-based correction. Each of 10,000 replicates sampled 23 wards with replacement and moved every block and reconstructed conditional density of a selected ward together. This permits arbitrary dependence among blocks within a ward but does not adjust dependence across ward boundaries. The pointwise lower limit remained positive through $1.027 \times 10^4\,\mathrm{m}^2$, while the maximum-studentized simultaneous lower limit remained positive through $9.152 \times 10^3\,\mathrm{m}^2$. At $10^4\,\mathrm{m}^2$, the pointwise interval was [0.0185, 0.450], 98.57% of replicates were positive, and the full-grid simultaneous interval was $[-0.0654, 0.558]$.

S5 Robustness of the curvature regime

Table S1 separates the well-supported lower-area result from the less stable higher-area crossover. Intervals are pointwise stratified-bootstrap intervals.

Across 69 bandwidth–ward-deletion combinations, $\mathcal{B}(10^4)$ ranged from 0.049 to 0.686, whereas $\mathcal{B}(2 \times 10^4)$ ranged from -0.892 to 0.147. Equal-ward and retained-area weights preserved positive balance at the lower scale and negative balance at the higher scale. The three weight schemes gave $\Xi = 0.374$–0.803 at 10^4 and 1.347–2.060 at $2 \times 10^4\,\mathrm{m}^2$.

The independent spline estimator used stabilized log-density bins, cubic smoothing splines, generalized cross-validation, and numerical renormalization [2, 3]. Its first crossover was 1.51–1.64 $\times 10^4\,\mathrm{m}^2$ across bin widths 0.125, 0.15, and 0.175. At $10^4\,\mathrm{m}^2$,

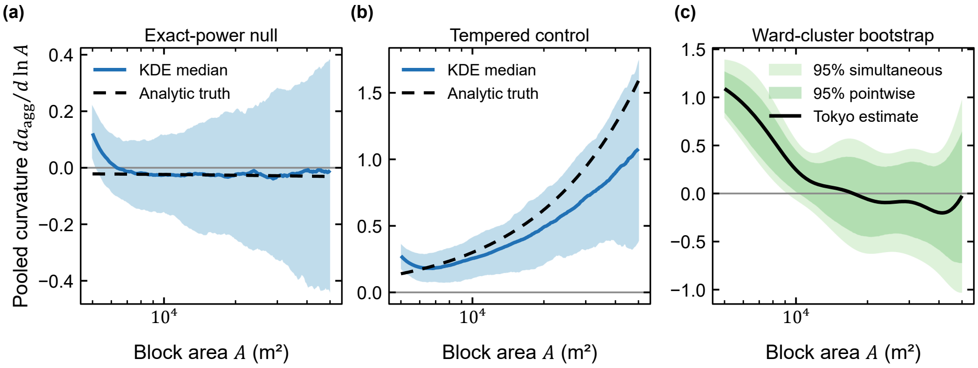

Fig. S1 Synthetic recovery and ward-cluster uncertainty. (a) Exact-power null: analytic pooled curvature, median reconstructed curvature, and central 95% simulation envelope across 500 datasets. (b) Tempered-power positive control with $\eta = 0.06$. (c) Empirical Tokyo curvature with pointwise and simultaneous 95% whole-ward cluster-bootstrap bands from 10,000 replicates. The synthetic panels use observed ward sample sizes and the empirical analysis settings.

Table S1 Selected dimensionless diagnostics for empirical block-count weights.

Quantity	$A \simeq 10^4\ \mathrm{m}^2$	$A \simeq 2 \times 10^4\ \mathrm{m}^2$
η_R	0.336 [0.224, 0.415]	0.308 [0.161, 0.397]
Ξ	0.374 [0.274, 0.610]	1.347 [0.57, 15.6]
$\mathcal{B}$	0.455 [0.242, 0.570]	−0.148 [−1.000, 0.277]

its within-ward term 0.471 exceeded the variance 0.160. Four common lower thresholds from 1850 to 3000 m^2 preserved the lower-area sign and gave kernel crossovers near $1.74 \times 10^4\ \mathrm{m}^2$.

S6 Parametric and held-out comparisons

Table S2 reports absolute pooled fits. Every family was rejected despite a clear relative ranking.

Table S2 Pooled common-support model comparison. KS p uses 1000 simulated datasets with refitting.

Model	ΔAIC	ΔBIC	KS p
Pareto	1635.8	1627.0	0.0010
Cutoff Pareto	1001.6	1001.6	0.0010
Stretched exponential	220.6	220.6	0.0010
Lognormal	0.0	0.0	0.0010

For predictive falsification, each of 50 prespecified splits independently held out 30% of blocks in every ward. Models were fitted only to training blocks. Held-out observations determined log predictive density; a reflected kernel density estimate of held-out blocks supplied curvature, crossover, and RDN-sorting targets. Table S3 gives the main summaries. Log-predictive-density gain is a mean relative to Pareto; the remaining entries are medians.

Table S3 Held-out predictive summaries over 50 within-ward splits.

Model	LPD gain per block	Curvature RMSE	Crossover log error	RDN curve RMSE	RDN shift error
Pareto	0	0.476	2.303	0.339	0.736
Cutoff Pareto	0.010	0.538	2.303	0.143	**0.162**
Stretched exponential	0.020	0.454	1.132	0.114	0.165
Lognormal	**0.023**	**0.367**	**0.896**	**0.104**	0.243

No family was best for every held-out feature. These comparisons rank finite-range approximations but do not identify a microscopic process.

References

[1] Bingham, N. H., Goldie, C. M. & Teugels, J. L. *Regular Variation* (Cambridge University Press, Cambridge, 1989).

[2] Silverman, B. W. On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Stat.* **10**, 795–810 (1982).

[3] Kooperberg, C. & Stone, C. J. A study of logspline density estimation. *Comput. Stat. Data Anal.* **12**, 327–347 (1991).