

Supplementary Material

A Distinct Form, Not a Distinct Divide: Independent Identification, Localized Price Discounts, and Within-Group Service Inequality of Urban Villages in Shenzhen

This supplement reports reproducibility checks and fuller estimates referenced in the main manuscript. All housing models use the common 8,400-observation district-fixed-effect sample unless noted otherwise.

Classification and sample construction

Table S1: Classification validation and grid-scale sensitivity.

Population or scale	n	Positive share	ROC AUC	PR AUC	Normalized PR gain
All 500 m cells	6,386	0.050	0.911	0.222	0.181
Village and formal cells	1,532	0.208	0.955	0.856	0.818
250 m grid	24,570	0.058	0.916	—	—
1,000 m grid	1,710	0.037	0.898	—	—

Table S2: Selected classification-threshold sensitivity.

Land overlap	Point buffer	Village cells	Prevalence	PR AUC	Normalized gain
0.20	250 m	431	0.067	0.269	0.216
0.30	250 m	318	0.050	0.222	0.181
0.50	250 m	237	0.037	0.161	0.129
0.30	500 m	789	0.124	0.338	0.245
0.30	750 m	1,521	0.238	0.436	0.260
0.30	1,000 m	2,308	0.361	0.522	0.251

Table S3: Housing and cell sample flow.

Step	Observations
Original housing communities	9,544
Missing price removed	1,000
Implausible price removed	4
Incomplete covariates removed	138
Regression sample before district screen	8,402
Missing district removed	2
Common housing-model sample	8,400
Duplicate-coordinate excess records in common sample	32
Baseline grid cells	6,386
Built cells	5,766
Residential Theil sample	1,505
Zero-service cells retained	14

Table S4: WorldPop count audit. The **ppp** raster is interpreted as people per pixel.

Quantity	Population
Source raster sum	15,198,640
Reprojected 30 m label-grid extent	15,159,948
Reprojection mass ratio	0.9975
Retained built-up analysis cells	14,861,407
Official contextual population	17,600,000
Source-to-official ratio	0.8636

Population mass balance

Housing-model robustness

Table S5: Buffer-union coefficient across covariate specifications.

Specification	Coefficient	95% CI	R^2	Maximum VIF
Parsimonious	0.0703	[0.0647, 0.0759]	0.534	7.6
Compact	0.0687	[0.0632, 0.0742]	0.548	46.1
Saturated	0.0572	[0.0515, 0.0628]	0.566	185.3

Table S6: Saturated HC3 coefficient estimates. Individual control coefficients are not interpreted because of multicollinearity.

Term	Coefficient	SE	CI low	CI high
Intercept	10.0789	0.2578	9.5735	10.5842
Nanshan district	0.7932	0.0378	0.7192	0.8672
Pingshan district	-0.1336	0.0511	-0.2337	-0.0335
Bao'an district	0.2241	0.0334	0.1587	0.2895
Yantian district	0.4396	0.0473	0.3468	0.5323
Futian district	0.3895	0.0540	0.2837	0.4952
Luohu district	0.3991	0.0422	0.3163	0.4819
Longhua district	0.0571	0.0372	-0.0158	0.1300
Longgang district	0.0414	0.0350	-0.0273	0.1101
Log CBD distance	-0.1671	0.0190	-0.2043	-0.1299
Log buffer-union distance	0.0572	0.0029	0.0515	0.0628
Log floor area ratio	0.3445	0.0276	0.2904	0.3987
Mean height	0.0034	0.0008	0.0018	0.0049
Footprint ratio	-1.0195	0.1430	-1.2997	-0.7393
Service points	0.0018	0.0003	0.0012	0.0024
Vitality points	-0.00014	0.00005	-0.00024	-0.00004
NDVI	0.7819	0.1041	0.5779	0.9859
Blue-green coverage	0.6839	0.0588	0.5686	0.7993
Park ratio	-0.2285	0.0493	-0.3252	-0.1318
Road density	0.4036	0.2865	-0.1578	0.9651
Log population density	0.0148	0.0067	0.0017	0.0280
Nighttime light	0.00100	0.00014	0.00074	0.00127
Land-use entropy	-0.1522	0.0378	-0.2262	-0.0781

Table S7: Variables with elevated VIF in compact and saturated specifications.

Variable	Compact VIF	Saturated VIF
CBD distance	46.1	185.3
Population density	—	125.6
Nighttime light	—	52.6
NDVI	29.1	41.2
Footprint ratio	—	36.8
Floor area ratio	9.8	30.7
Blue-green coverage	10.3	26.2
Land-use entropy	—	16.4

Table S8: Duplicate-coordinate sensitivity.

Estimate	All 8,400	Unique 8,368	Difference
Parsimonious HC3 coefficient	0.07028	0.07016	-0.00012
KNN 8 SAR direct impact	0.04578	0.04564	-0.00014
KNN 8 SEM coefficient	0.07403	0.07353	-0.00050

Table S9: Raw KNN 8 SAR coefficients for distance bands; main-text Table 6 reports multiplier-based impacts.

Band	Raw coefficient	CI low	CI high
0–250 m	-0.120	-0.188	-0.052
250–500 m	-0.021	-0.090	0.048
500–1,000 m	-0.018	-0.086	0.050
1–2 km	0.014	-0.059	0.086

Spatial-block sensitivity

Table S10: Road-density spatial-block sensitivity.

Block size	Blocks	Mixed-class blocks	BH-adjusted p	Interpretation
1 km	749	155	0.033	Significant
2 km	307	145	0.039	Marginal
4 km	110	88	0.053	Not significant

Reproducibility and data-availability files

The separately submitted package contains open derived results, processed analysis data requiring redistribution review, reproduction code, and source documentation. No raw provider-licensed or platform-derived records are included. Key files corresponding directly to supplementary tables are:

- A2_classification_validation_extended.csv
- A3_threshold_2d_sensitivity.csv
- A4_grid_scale_sensitivity.csv
- A5_spatial_block_size_sensitivity.csv
- C_housing_sample_accounting.csv
- C6_full_ols_covariates.csv
- C8_vif.csv and C8_parsimonious_vif.csv
- C10_distance_bands_spatial.csv
- C11_compact_spatial_models.csv
- C13_duplicate_coordinate_sensitivity.csv
- C14_duplicate_coordinate_spatial_sensitivity.csv
- E1_population_mass_balance.csv

The historical word “compact” in the C11 filename is retained for code compatibility; its current contents are the principal parsimonious spatial models.