Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Two-Scale Housing-Price Analysis Using Spatial Autoregression and Multiscale Geographically Weighted Regression

Wanqiu Zhang
Henan University of Technology
Cunchang Gu
gucunchang@163.com
Henan University of Technology
Zhouying Wu
Henan University of Technology

Article

Keywords:

Posted Date: November 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11584124/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Two-Scale Housing-Price Analysis Using Spatial Autoregression and Multiscale Geographically Weighted Regression

Wanqiu Zhang[1], Cunchang Gu[1*], Zhouying Wu[1]

[1]Henan University of Technology, China.

*Corresponding author(s). E-mail(s): gucunchang@163.com;
Contributing authors: 19838757197@163.com; 17516815061@163.com;

Abstract

Housing datasets often contain several dwelling types within the same residential community, creating a mismatch between record-specific attributes and shared spatial conditions. This study develops a two-scale framework that links dwelling-type and community-level datasets through consistently constructed indicators. At the dwelling-type scale, ordinary least squares models retain individual housing attributes and use community-clustered inference and community-grouped cross-validation. At the community scale, spatial econometric models examine spatial dependence and multiplier associations, while geographically weighted regression and multiscale geographically weighted regression evaluate spatially varying relationships and variable-specific scales. An application to 461 dwelling-type records across 198 communities in Zhengzhou identifies pronounced spatial dependence in community prices and supports a spatial autoregressive specification. Floor area, metro accessibility, and opening-year area-level GDP have statistically supported spatial effects. The multiscale results distinguish a localized shopping-centre relationship from broader metro-accessibility and near-global GDP relationships. Sensitivity checks cover alternative spatial weights, community price aggregation, and minimum bandwidth settings. By preserving within-community housing variation while defining spatial relationships at the community level, the framework provides a consistent and transferable basis for analysing housing-price relationships and their spatial scales across urban contexts.

Keywords: housing prices, spatial dependence, spatial econometrics, multiscale geographically weighted regression

1 Introduction

Housing-price analysis requires a consistent connection between the characteristics being measured, the units being observed, and the spatial relationships being modelled. The hedonic approach relates housing prices to structural, neighbourhood, accessibility, and economic attributes (Rosen 1974; Hill 2013). These attributes are observed at different levels: floor area and dwelling configuration describe individual housing records, while community facilities, geographic location, and surrounding amenities are often shared by several records. Nearby communities may also share locational conditions and market environments, producing spatial dependence documented across different housing markets (Anselin and Lozano-Gracia 2009; Osland 2010; Bourassa et al. 2010; Gröbel and Thomschke 2018). A coherent empirical design must therefore address both the organization of housing information and the spatial structure of housing-price relationships.

This requirement is particularly important for datasets containing multiple dwelling types within residential communities. Aggregating all records to community averages removes the dwelling-type variation available for baseline analysis. Conversely, constructing spatial neighbourhoods directly from repeated records at shared locations makes the representation of a community depend on the number of dwelling types recorded for it. Hierarchical approaches to geographically weighted housing analysis also highlight the importance of distinguishing levels within housing data (Hu et al. 2022). These considerations motivate linked datasets in which dwelling-type observations retain their individual attributes and community-level observations provide the geographic units for spatial analysis. The central design question is how to construct these datasets and their indicators consistently while assigning each scale a clear analytical purpose.

Accessibility further illustrates why data construction and spatial modelling need to be considered together. Research on transport, education, and green space documents

housing-price relationships associated with different dimensions of urban accessibility (Karlsson 2011; Yang et al. 2020; Debrezion et al. 2007; Wen et al. 2018; Liu et al. 2024). Other studies examine how these relationships vary across price segments, transport contexts, locations, and development periods (Jin et al. 2022; Yang et al. 2022; Zhou et al. 2022). Multiscale housing studies additionally identify differences in the spatial patterns of accessibility and other explanatory variables (Liu et al. 2022; Sachdeva et al. 2022; Lu et al. 2023). Accessibility indicators can therefore share a common observational unit while exhibiting different geographic patterns in their associations with prices. A framework that distinguishes observational scale from relationship-specific spatial scale can accommodate both features.

Spatial econometric and geographically weighted models address complementary aspects of this problem. Spatial econometric models characterize dependence among geographically connected observations and, where supported by the selected specification, spatial multiplier associations. Geographically weighted regression (GWR) allows coefficients to vary across locations (Brunsdon et al. 1996; Bitter et al. 2007). Multiscale geographically weighted regression (MGWR) extends this approach by estimating a separate bandwidth for each coefficient surface (Fotheringham et al. 2017; Oshan et al. 2019), with applications to spatially differentiated housing markets (Chen et al. 2024). Variable-specific bandwidths provide information on differences in the spatial scales of estimated relationships (Fotheringham et al. 2022). Combining these analytical perspectives within a consistent data structure makes it possible to examine dependence among community prices alongside variation in individual housing-price relationships.

The study contributes to housing-price analysis in three respects: (i) it establishes a linked two-scale data and indicator structure that preserves dwelling-type information while defining spatial relationships among communities; (ii) it assigns complementary roles to baseline, spatial econometric, and multiscale models, connecting each method to a distinct analytical question; and (iii) it integrates specification assessment and

sensitivity checks to examine how the findings depend on data aggregation, spatial connectivity, and local estimation settings. The application to Zhengzhou illustrates the construction of the linked indicators and the complementary evidence obtained at the two observational scales.

The remainder of the paper is organized as follows. Section 2 presents the empirical framework, model specifications, robustness procedures, and implementation. Section 3 applies the framework to Zhengzhou and reports the empirical results. Section 4 discusses the substantive findings and policy implications, and Section 5 concludes.

2 Empirical Methodology

2.1 Two-Scale Empirical Framework

Housing datasets often contain several dwelling types within the same residential community. Dwelling characteristics can vary across these records, while their location, accessibility, and community attributes are shared. An analysis based entirely on community averages would discard dwelling-type variation. Conversely, defining spatial neighbours directly among all dwelling-type records would give greater representation to communities with more recorded types. These features motivate a two-scale framework that retains dwelling-type observations for baseline modelling and constructs one observation per community for spatial analysis.

The framework requires housing records linked to community identifiers, consistently defined unit prices, dwelling and community attributes, geographic coordinates, and relevant accessibility and economic measures. Data preparation checks record duplication, missing information, measurement units, and the consistency of attributes shared within communities. Economic indicators are matched to the appropriate administrative area and reference year. The meaning and timing of the price measure are documented so that the resulting associations can be interpreted consistently.

Let $i = 1, \ldots, n$ index communities and $r = 1, \ldots, m_i$ index the dwelling types recorded in community i. The dwelling-type dataset contains $N = \sum_{i=1}^{n} m_i$ observations and preserves their individual prices and attributes. The community dataset contains n observations, with the principal outcome defined as

$$y_i = \ln[\text{median}_{r=1,\ldots,m_i}(P_{ir})]\,,$$

where $P_{ir} > 0$ denotes the observed unit price. A community with a single record retains that price. Record-specific floor area is also summarized by its community median, while shared community attributes retain their original values. Aggregation is based on community identifiers, so distinct communities can remain separate even when their recorded coordinates coincide.

The two datasets serve complementary purposes. Dwelling-type modelling estimates average conditional associations and evaluates prediction for held-out communities. Community-level spatial econometric modelling examines dependence and spatial multiplier associations, while geographically weighted modelling examines local coefficients and relationship-specific spatial scales. The two observational scales are distinct from the variable-specific bandwidths estimated within the community-level analysis.

2.2 Models at the Two Observational Scales

2.2.1 Dwelling-Type Baseline Model

The dwelling-type baseline follows the hedonic approach, relating housing prices to observable attributes (Rosen 1974; Hill 2013). Its general specification is

$$\ln P_{ir} = \alpha_d + \mathbf{h}_{ir}'\boldsymbol{b}_h + \mathbf{c}_i'\boldsymbol{b}_c + \mathbf{d}_i'\boldsymbol{b}_d + \mathbf{e}_i'\boldsymbol{b}_e + \mathbf{t}_i'\boldsymbol{\delta} + \varepsilon_{ir}, \tag{1}$$

where α_d is the intercept. The column vectors $\mathbf{h}_{ir}$, $\mathbf{c}_i$, $\mathbf{d}_i$, and $\mathbf{e}_i$ contain dwelling characteristics, shared community attributes, accessibility measures, and area-level economic conditions, respectively. Economic conditions are matched to the administrative area and project opening year associated with community i. The vector $\mathbf{t}_i$ contains opening-year indicators with one reference category omitted. The vectors $\boldsymbol{b}_h$, $\boldsymbol{b}_c$, $\boldsymbol{b}_d$, $\boldsymbol{b}_e$, and $\boldsymbol{\delta}$ contain the corresponding coefficients; ε_{ir} is a scalar disturbance and the prime denotes transpose. Distance-based accessibility variables are expressed as $\ln(1 + D)$, where D is a nonnegative distance measured in a common unit.

The model is estimated by ordinary least squares (OLS), with standard errors clustered by residential community. This allows heteroskedasticity and correlated disturbances among records within the same community. The pooled specification uses variation both within and between communities; community clustering adjusts inference for the repeated observations but does not remove unobserved differences between communities.

Candidate specifications vary selected housing or economic covariates while retaining a common set of substantive controls. Selection combines parsimony with five-fold cross-validation grouped by community. All records from a community remain in the same fold. Predictive performance is assessed using log-scale root mean squared error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2), together with raw-scale mean absolute percentage error (MAPE). These measures evaluate prediction for records from communities excluded from the corresponding estimation fold. The selected baseline informs the covariate specification for the community-level analysis.

2.2.2 Community-Level Spatial Models

1) Spatial weights and dependence diagnostics. Community coordinates recorded in the World Geodetic System 1984 (WGS84) are transformed to a local azimuthal equidistant projection centred on the study city. Inter-community Euclidean distances

are then calculated in metres using the projected coordinates. Spatial connectivity is represented by a K-nearest-neighbour (KNN) matrix. For communities i and j,

$$w_{ij}^{*} = \begin{cases} 1, & j \in \mathcal{N}_K(i), \\ 0, & \text{otherwise}, \end{cases} \qquad w_{ii}^{*} = 0, \tag{2}$$

where $\mathcal{N}_K(i)$ contains the K nearest other communities, with $1 \leq K < n$. Row standardization gives

$$w_{ij} = \frac{w_{ij}^{*}}{\sum_{\ell=1}^{n} w_{i\ell}^{*}}, \qquad \mathbf{W} = (w_{ij})_{n \times n}. \tag{3}$$

This construction assigns neighbours to each community despite uneven sampling density. The choice of K defines the spatial structure being examined and is assessed through sensitivity analysis (Gerkman and Ahlgren 2014).

Global spatial autocorrelation is measured by Moran's I:

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \qquad S_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}, \tag{4}$$

where $\bar{y}$ is the sample mean of the community-level log prices and S_0 is the sum of the spatial weights. Significance is assessed using 9,999 random permutations. The same statistic is calculated for model residuals by replacing y_i with the corresponding residual.

An OLS model with the community-level covariates provides the non-spatial reference. Standard and robust Lagrange multiplier (LM) tests examine remaining spatial lag and spatial error dependence. Model comparison then considers the spatial autoregressive (SAR), spatial error (SEM), and spatial Durbin (SDM) models alongside these diagnostics (Florax et al. 2003).

2) Spatial specifications and model selection. The SDM is written as

$$(\mathbf{I}_n - \rho\mathbf{W})\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}_c\boldsymbol{\theta} + \mathbf{T}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}. \tag{5}$$

where n is the number of residential communities and $\mathbf{y}$ contains their log housing prices. The matrix $\mathbf{W}$ describes the spatial relationships among communities, and ρ measures the strength of spatial dependence in housing prices. The matrix $\mathbf{I}_n$ is the identity matrix.

The matrix $\mathbf{X}$ contains a constant term and the housing, community, accessibility, and economic covariates, while $\boldsymbol{\beta}$ contains the intercept and their corresponding coefficients. The matrix $\mathbf{X}_c$ contains the covariates whose spatially lagged values are included, so $\mathbf{W}\mathbf{X}_c$ represents the corresponding characteristics of neighbouring communities and $\boldsymbol{\theta}$ contains their coefficients. The matrix $\mathbf{T}$ contains the opening-year indicators, with one year omitted as the reference category, and $\boldsymbol{\gamma}$ contains their coefficients. The year indicators are not spatially lagged. Finally, $\boldsymbol{\varepsilon}$ represents the unobserved disturbances.

Imposing $H_0 : \boldsymbol{\theta} = \mathbf{0}$ yields the SAR model:

$$(\mathbf{I}_n - \rho\mathbf{W})\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{T}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}. \tag{6}$$

The quantity $(\mathbf{W}\mathbf{y})_i = \sum_{j=1}^{n} w_{ij}y_j$ is the weighted mean of neighbouring log prices, so the model represents a jointly determined system of community prices. Likelihood-ratio (LR) and Wald tests evaluate the SDM-to-SAR restriction (Koley and Bera 2024). Failure to reject is considered alongside the other diagnostics when assessing the more parsimonious specification.

The SEM represents dependence in the disturbances:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{T}\boldsymbol{\gamma} + \mathbf{u}, \\ \mathbf{u} &= \lambda \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}, \end{aligned} \tag{7}$$

where $\mathbf{u}$ is the $n \times 1$ spatially correlated disturbance vector and λ is its spatial dependence parameter. Spatial models are estimated by maximum likelihood, with spatial parameters in the admissible domain of the relevant spatial multiplier. Selection jointly considers LM diagnostics, restriction tests, the Akaike information criterion (AIC), the Bayesian information criterion (BIC), parsimony, and residual Moran's I. The preferred specification is determined by the evidence in the application.

3) Spatial multiplier effects. When the SAR specification is supported, its reduced form is

$$\mathbf{y} = (\mathbf{I}_n - \rho \mathbf{W})^{-1}(\mathbf{X}\boldsymbol{\beta} + \mathbf{T}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}). \tag{8}$$

For covariate k, let $\mathbf{x}_k$ denote the corresponding column of $\mathbf{X}$. Holding $\mathbf{W}$ and other covariates fixed, the matrix of partial derivatives of fitted conditional means is

$$\mathbf{S}_k = \frac{\partial E(\mathbf{y} \mid \mathbf{X}, \mathbf{T})}{\partial \mathbf{x}_k'} = \beta_k (\mathbf{I}_n - \rho \mathbf{W})^{-1}. \tag{9}$$

Its (i, j) element is the partial derivative of the expected log price in community i with respect to covariate k in community j.

$$\mathrm{ADE}_k = \frac{1}{n} \operatorname{tr}(\mathbf{S}_k), \tag{10}$$

$$\mathrm{ATE}_k = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} (\mathbf{S}_k)_{ij}, \tag{11}$$

$$\mathrm{AIE}_k = \mathrm{ATE}_k - \mathrm{ADE}_k. \tag{12}$$

Here, $\mathrm{tr}(\cdot)$ denotes the matrix trace. The average direct effect (ADE) averages the diagonal elements of $\mathbf{S}_k$, the average total effect (ATE) averages its row sums, and the average indirect effect (AIE) is their difference.

4) Spatially varying relationships and bandwidths. GWR and MGWR examine variation in conditional associations across community locations. The outcome and selected covariates are standardized before estimation:

$$z_{ik} = \frac{x_{ik} - \bar{x}_k}{s_k}, \qquad \widetilde{y}_i = \frac{y_i - \bar{y}}{s_y}, \tag{13}$$

Here, $k = 1, \ldots, p_G$ indexes the covariates selected for the local models, and x_{ik} denotes their values before standardization, after any distance transformation. The quantities $\bar{x}_k$ and s_k are its sample mean and standard deviation, and $\bar{y}$ and s_y are those of the log-price outcome. All variables being standardized must have positive sample variance.

For community i at projected coordinates (u_i, v_i), GWR is specified as

$$\widetilde{y}_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p_G} \beta_k(u_i, v_i) z_{ik} + \eta_i, \tag{14}$$

where p_G is the number of local-model covariates, $\beta_0(u_i, v_i)$ is the local intercept, $\beta_k(u_i, v_i)$ is a local coefficient, and η_i is the disturbance of the standardized model. GWR estimates all coefficient surfaces using a common bandwidth (Brunsdon et al. 1996; Bitter et al. 2007). MGWR allows a separate bandwidth for each surface:

$$\widetilde{y}_i = \beta_{0,bw_0}(u_i, v_i) + \sum_{k=1}^{p_G} \beta_{k,bw_k}(u_i, v_i) z_{ik} + \eta_i, \tag{15}$$

where bw_0 and bw_k are adaptive bandwidths for the intercept and covariate k, expressed as numbers of nearest communities (Fotheringham et al. 2017; Oshan et al. 2019). Small bandwidths indicate more localized estimation, while bandwidths approaching

the sample size indicate broader or near-global relationships. These describe the spatial scale of coefficient estimation, with their geographic extent also depending on the distribution of sampled communities.

Both models use an adaptive bi-square kernel. Bandwidth selection minimizes the corrected Akaike information criterion (AICc), using iterative multibandwidth calibration for MGWR. The local covariate set is assessed for collinearity, and any differences from the global specification are reported in the empirical application. Model assessment combines AICc, effective parameter counts, R^2, adjusted R^2, residual spatial autocorrelation, and local condition numbers (Kang and Oshan 2025). Local inference uses variable-specific adjusted critical values for multiple testing at a nominal 5% significance level. Bandwidths and the proportions of statistically supported local coefficients are reported separately: bandwidth describes estimation scale, while significance describes statistical support. Spatial econometric and geographically weighted models consequently provide complementary evidence on dependence and heterogeneity.

2.3 Robustness and Sensitivity Assessment

Robustness is assessed by varying spatial connectivity, the construction of community prices, and the admissible scales of local estimation. Conclusions are evaluated from the stability of coefficient directions, effect magnitudes and uncertainty, specification tests, and residual diagnostics across these changes.

Spatial-weight sensitivity is examined by varying K in the binary KNN matrix and replacing equal neighbour weights with inverse-distance weights. All matrices remain row-standardized. For distinct communities sharing coordinates, a zero distance is replaced by the minimum positive observed inter-community distance solely when calculating inverse-distance weights; this convention avoids undefined weights while retaining the communities as separate observations and is treated as part of the weight-definition sensitivity analysis. The recorded coordinates are retained. Moran's

I, spatial-model diagnostics, restriction tests, and the principal effect estimates are compared across the alternative matrices. This assesses whether the dependence evidence and substantive associations persist under different definitions of spatial connectivity.

Aggregation sensitivity is assessed by replacing the community median price with the arithmetic mean of observed dwelling-type prices and then taking its logarithm. The selected spatial specification is re-estimated with the same covariates and weights. Comparisons examine the spatial parameter, coefficient directions, prediction errors, and residual spatial autocorrelation. This isolates the influence of the community price summary used to construct the dependent variable.

MGWR sensitivity is examined by reducing the minimum admissible adaptive bandwidth and repeating calibration. The resulting bandwidths, local coefficient surfaces, AICc, effective parameter counts, and collinearity diagnostics are compared. Similar coefficient patterns across numerically stable solutions support the interpretation of the main spatial scales. A smaller AICc accompanied by severe ill-conditioning is reported as a numerical sensitivity result. The neighbourhood sizes and bandwidth bounds used in these comparisons are specified in the empirical application.

2.4 Software and Implementation

The framework is implemented in Python, using Python 3.14 for the present analysis. Dwelling-type OLS models and community-clustered standard errors are estimated with `statsmodels`, while `scikit-learn` provides community-grouped cross-validation through `GroupKFold`. Coordinate transformation uses `pyproj`. Spatial weights and Moran's I are calculated with `libpysal` and `esda`, respectively. Maximum-likelihood estimation of the spatial models uses `spreg`, and GWR, MGWR, and adaptive bandwidth selection use `mgwr`.

3 Empirical Results and Analysis

This section applies the two-scale empirical framework to Zhengzhou, the capital of Henan Province in central China. The sample covers 13 administrative or functional areas and includes both the densely developed urban core and more peripheral housing-market environments. Figure 1 shows the spatial distribution of the sampled residential communities.

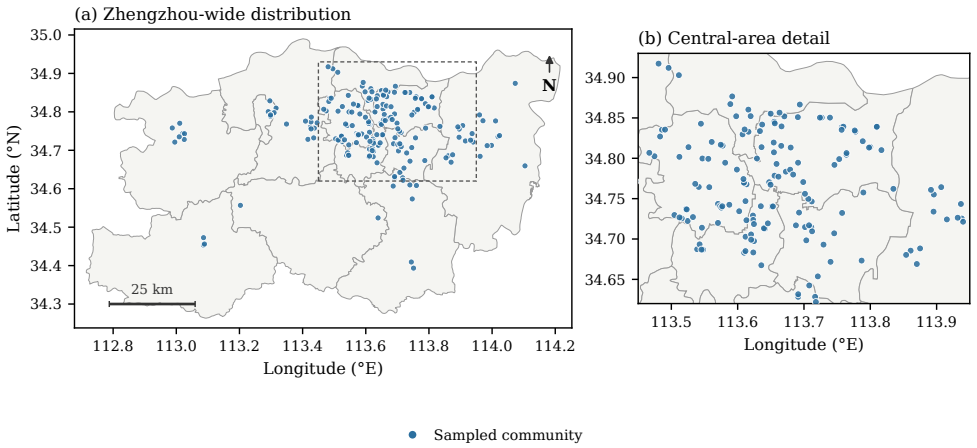


Fig. 1 Location of sampled residential communities in Zhengzhou
Notes: Panel (a) shows the Zhengzhou-wide distribution of the housing sample, and panel (b) enlarges the central area indicated by the dashed rectangle. Points represent sampled residential communities.

Housing information was collected in 2026 from the Zhenghaofang housing-information platform. The price variable is the filing unit price displayed on the platform rather than an observed transaction price. The collection covered all accessible residential developments that met the study's predefined eligibility criteria, and screening identified no missing values or duplicate records. The final sample contains 461 dwelling-type records nested within 198 communities. These records form two linked datasets: the dwelling-type dataset retains the price and attributes of each recorded dwelling type, while the community-level dataset contains one observation per community and uses the median filing unit price as its principal outcome. The corresponding community

mean is used for robustness analysis. Because platform accessibility and the eligibility criteria delimit the sample, the data should not be interpreted as a probability sample of all Zhengzhou housing. POI coordinates for metro stations, parks, shopping centres, hospitals, and schools were obtained from Amap. In a two-step QGIS workflow, the Amap GCJ-02 coordinates were first converted to WGS84 and were then projected to a metric coordinate system; straight-line distances from each residential community to the nearest facility in each category were subsequently calculated in metres.

The 198 communities correspond to 195 unique coordinate pairs because three pairs of adjacent communities share coordinates; they remain separate because their housing information differs. For the Zhengzhou application, a local azimuthal equidistant projection centred on the city is used to calculate inter-community Euclidean distances in metres. The analysis uses log unit price, dwelling and community characteristics, five log-transformed accessibility distances, opening-year area-level GDP, and opening-year indicators. Table 1 defines these variables and their transformations. The results below first use the dwelling-type data to select and estimate the baseline specification, and then use the community data to examine spatial dependence, multiplier effects, and variable-specific spatial scales.

3.1 Dwelling-Type Model Selection and Baseline Results

The dwelling-type dataset contains 461 observations from 198 residential communities. Displayed filing unit prices average CNY 13,962.44/m^2, with a median of CNY 13,497.16/m^2 and a range from CNY 4,195.01/m^2 to CNY 49,400.22/m^2. The raw distribution is right-skewed (skewness = 1.642), whereas log transformation produces a substantially more symmetric distribution (skewness = −0.093). Mean dwelling floor area is 124.70 m^2 and the median is 120.96 m^2. Accessibility also varies across the sample: the median straight-line distances to the nearest metro station, park, shopping

Table 1 Principal variables and data construction

Variable	Analytical level	Definition and transformation
Unit price	Dwelling type / community	Displayed residential unit price (CNY/m^2). The natural logarithm is used; the community median is the principal spatial outcome.
Floor area	Dwelling type / community	Dwelling floor area (m^2). Record-specific values are used in the baseline model and the community median in the principal spatial analysis.
FAR	Community	Floor area ratio, rescaled from its representation in the source data to a conventional numerical scale.
Green-space ratio	Community	Reported proportion of green space within the residential community.
Gas and central-heating availability	Community	Binary indicator equal to one when both natural gas and central heating are available and zero otherwise; communities with natural gas only are coded as zero.
Metro distance	Community	Straight-line distance to the nearest metro station (m), transformed as $\ln(1 + D)$.
Park distance	Community	Straight-line distance to the nearest park (m), transformed as $\ln(1 + D)$.
Shopping-centre distance	Community	Straight-line distance to the nearest shopping centre (m), transformed as $\ln(1 + D)$.
Hospital distance	Community	Straight-line distance to the nearest hospital (m), transformed as $\ln(1 + D)$; excluded from the final MGWR specification following collinearity diagnostics.
School distance	Community	Straight-line distance to the nearest school (m), transformed as $\ln(1 + D)$.
GDP	Area × opening year	Area-level GDP (CNY 100 million), matched by administrative area and project opening year and interpreted as opening-year area-level economic context.
Opening year	Project	Project opening year, represented by fixed effects in the dwelling-type baseline and spatial econometric models.

Notes: A larger transformed distance represents poorer accessibility. Hospital distance and opening-year fixed effects are not included in the final MGWR specification.

centre, hospital, and school are approximately 1.34, 3.95, 5.86, 6.24, and 2.14 km, respectively.

Four candidate specifications are compared to determine the dwelling-type baseline used in the subsequent analysis. All four specifications include dwelling floor area, FAR, green-space ratio, gas and central-heating availability, the five accessibility measures, and opening-year fixed effects. B1 additionally includes opening-year area-level GDP.

B2 adds the dwelling-type score to B1. B3 replaces GDP with per-capita disposable income, while B4 includes both GDP and income.

Table 2 compares the four candidate specifications. Adding the dwelling-type score in B2 produces little improvement in in-sample fit and slightly weaker community-grouped prediction. Replacing GDP with per-capita disposable income in B3 substantially reduces both explanatory and predictive performance. Although B4 has a slightly lower in-sample AIC and CV MAE, B1 has a lower CV RMSE and a higher CV R^2. Given these small differences, the GDP-based B1 specification is retained on the combined grounds of predictive performance and parsimony.

Table 2 Comparison of dwelling-type baseline specifications

Specification	R^2	Adj. R^2	AIC	BIC	CV RMSE	CV MAE	CV R^2
B1: GDP	0.688	0.676	71.688	146.089	0.280	0.220	0.603
B2: GDP + dwelling score	0.690	0.677	71.208	149.699	0.282	0.221	0.600
B3: Income	0.617	0.602	166.497	240.898	0.307	0.240	0.521
B4: GDP + income	0.690	0.678	70.719	149.254	0.281	0.219	0.601

Notes: $N = 461$. The dependent variable is log displayed filing unit price. All specifications include opening-year fixed effects and use community-clustered standard errors. CV statistics are based on five-fold GroupKFold cross-validation by residential community.

Under five-fold community-grouped cross-validation, B1 has a mean log-scale RMSE of 0.280, an MAE of 0.220, and an R^2 of 0.603. Its fold-specific raw-scale MAPE ranges from 20.51% to 24.11%, with a mean of 22.28%. Because every dwelling-type record from a community remains in the same fold, these statistics evaluate predictions for dwelling-type observations belonging to communities excluded from the corresponding estimation fold.

The estimated B1 coefficients show positive associations between displayed filing unit price and both dwelling floor area and opening-year area-level GDP. The floor-area coefficient is 0.0024 ($p < 0.001$), while the GDP coefficient is 0.000321 ($p < 0.001$). FAR is negatively associated with price ($\beta = -0.0773$, $p = 0.002$), as is transformed shopping-centre distance ($\beta = -0.1690$, $p < 0.001$). Because a larger distance denotes poorer

accessibility, the latter result indicates that greater proximity to a shopping centre is associated with a higher displayed filing unit price, conditional on the other covariates. Gas and central-heating availability and metro distance are close to, but do not reach, the conventional 5% significance threshold ($p = 0.061$ and $p = 0.054$, respectively); the remaining community and accessibility variables are not statistically distinguishable from zero in this specification. These estimates describe average conditional relationships at the dwelling-type level. Full coefficient estimates for the preferred dwelling-type specification are reported in Table 3.

Table 3 Full estimates from the preferred dwelling-type baseline model

Variable	Coefficient	Clustered SE	p-value	95% CI
Intercept	11.381970	0.486407	< 0.001	[10.428630, 12.335310]
Floor area	0.002352	0.000440	< 0.001	[0.001490, 0.003215]
FAR	−0.077327	0.024677	0.002	[−0.125692, −0.028962]
Green-space ratio	0.248960	0.334538	0.457	[−0.406722, 0.904643]
Gas and central-heating availability	0.122419	0.065421	0.061	[−0.005805, 0.250642]
Metro distance	−0.066356	0.034481	0.054	[−0.133938, 0.001226]
Park distance	−0.031475	0.021181	0.137	[−0.072990, 0.010039]
Shopping-centre distance	−0.168983	0.037799	< 0.001	[−0.243068, −0.094898]
Hospital distance	−0.033916	0.035410	0.338	[−0.103318, 0.035486]
School distance	−0.005417	0.024539	0.825	[−0.053513, 0.042679]
Opening-year area-level GDP	0.000321	0.000035	< 0.001	[0.000253, 0.000389]
Opening year: 2019	0.089929	0.081441	0.269	[−0.069693, 0.249551]
Opening year: 2020	0.079080	0.079045	0.317	[−0.075845, 0.234005]
Opening year: 2021	0.071003	0.085081	0.404	[−0.095752, 0.237758]
Opening year: 2022	0.006042	0.087804	0.945	[−0.166051, 0.178135]
Opening year: 2023	0.086456	0.081705	0.290	[−0.073683, 0.246595]
Opening year: 2024	0.129003	0.085009	0.129	[−0.037613, 0.295618]
Opening year: 2025	0.019931	0.086902	0.819	[−0.150393, 0.190255]

Notes: $N = 461$. The dependent variable is log displayed filing unit price. Standard errors are clustered by residential community. The omitted opening-year category is 2018. Distance variables are transformed as $\ln(1 + D)$.

The analysis next moves to the community-level dataset to examine whether housing prices remain spatially dependent after the observed characteristics are taken into account.

3.2 Community-Level Spatial Dependence and SAR Results

3.2.1 Spatial Dependence and Model Selection

Community housing prices exhibit pronounced positive spatial dependence under the baseline row-standardized binary KNN-5 matrix. Moran's I for log median displayed filing unit price is 0.7075 ($p = 0.0001$). After the observed housing, community, accessibility, economic, and opening-year characteristics are included in the non-spatial specification, residual spatial dependence remains detectable: the OLS residual Moran's I is 0.2555 ($p = 0.0001$). The corresponding Moran scatterplots are presented in Figure 2. The steep positive slope for community prices becomes smaller for the OLS residuals and nearly flat for the SAR residuals, tracing the reduction in spatial autocorrelation across the three stages of the analysis.

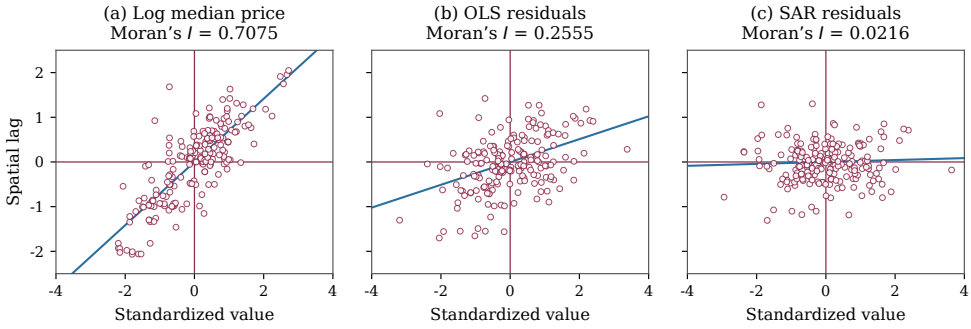


Fig. 2 Moran scatterplots for community housing prices and model residuals
Notes: The panels show (a) log median displayed filing unit price, (b) OLS residuals, and (c) SAR residuals under the row-standardized binary KNN-5 spatial weights matrix. Moran's I is 0.7075 for log median price, 0.2555 for OLS residuals, and 0.0216 for SAR residuals.

The LM diagnostics favour a spatial-lag specification over dependence confined to the error process. Although both conventional LM tests are significant, the robust LM-lag statistic remains significant after accounting for the competing alternative (15.330, $p < 0.001$), whereas the robust LM-error statistic does not (2.415, $p = 0.120$). The more general SDM also yields a positive spatial autoregressive coefficient, but neither

the LR test ($p = 0.469$) nor the Wald test ($p = 0.442$) rejects the joint restriction that the spatially lagged substantive covariates can be excluded. These results support reduction of the SDM to the more parsimonious SAR specification.

The model comparison in Table 4 provides additional evidence. The OLS model has an R^2 of 0.727 and an AIC of 40.012, but its residual spatial autocorrelation remains significant. The SAR estimate of ρ is 0.4382 ($p < 0.001$), and its AIC falls to 0.821. The SEM also contains a significant spatial parameter ($\lambda = 0.5807$, $p < 0.001$), while the SDM estimate of ρ is 0.4591 ($p < 0.001$). AIC favours the SAR specification, whereas BIC marginally favours the SEM. Model selection therefore draws on the robust LM diagnostics, the SDM restriction tests, parsimony, and residual behaviour in addition to the information criteria. Under the selected SAR model, residual Moran's I declines to 0.0216 and is not statistically significant, indicating no detectable residual spatial autocorrelation under the baseline weights matrix.

Table 4 Spatial-dependence diagnostics and global model comparison

Statistic or model	Estimate	Test statistic / AIC	p-value
Global Moran's I	0.7075	–	0.0001
OLS residual Moran's I	0.2555	–	0.0001
LM-lag	–	50.107	< 0.001
Robust LM-lag	–	15.330	< 0.001
LM-error	–	37.192	< 0.001
Robust LM-error	–	2.415	0.120
OLS	–	40.012	–
SAR (ρ)	0.4382	0.821	< 0.001
SEM (λ)	0.5807	3.402	< 0.001
SDM (ρ)	0.4591	11.136	< 0.001

Notes: $N = 198$. All spatial statistics use the row-standardized binary KNN-5 matrix. The third column reports the LM statistic for the diagnostic tests and AIC for the fitted models. The BIC values for OLS, SAR, SEM, and SDM are 99.201, 63.298, 62.591, and 106.496, respectively. The dependent variable is log median displayed filing unit price.

3.2.2 SAR Spatial Effects

Substantive interpretation of the SAR model is based on the effect decomposition in Table 5. Three variables have direct, indirect, and total effects whose simulation-based 95% confidence intervals exclude zero. Median dwelling floor area has a direct effect of 0.0046, an indirect effect of 0.0033, and a total effect of 0.0079. The direct estimate implies that a one-square-metre increase in median dwelling floor area is associated with an approximately 0.46% increase in community price through the average own-community effect, holding the other modelled characteristics constant.

Metro distance is transformed as $\ln(1 + D)$ to reduce right skewness and allow for a diminishing marginal association with housing prices. The transformed measure has negative direct, indirect, and total effects of -0.0522, -0.0371, and -0.0893, respectively. Their confidence intervals remain below zero, indicating that poorer metro accessibility is associated with lower displayed filing unit prices. Opening-year area-level GDP has positive direct, indirect, and total effects of 0.000150, 0.000106, and 0.000256, respectively, with all three confidence intervals above zero. This finding indicates a positive association between housing prices and the economic conditions of the administrative area at the time of project opening.

The confidence intervals for FAR, green-space ratio, gas and central-heating availability, and the remaining accessibility measures include zero. The global SAR estimates therefore do not provide sufficient evidence that their spatial effects are distinguishable from zero across the study area. For all variables, the indirect effects capture spatial multiplier associations under the specified spatial weights matrix and do not establish causal effects across communities.

Table 5 Direct, indirect, and total effects from the SAR model

Variable	Direct effect	Indirect effect	Total effect
Floor area	0.0046 [0.0034, 0.0058]	0.0033 [0.0018, 0.0055]	0.0079 [0.0056, 0.0108]
FAR	−0.0160 [−0.0553, 0.0216]	−0.0114 [−0.0440, 0.0163]	−0.0273 [−0.0976, 0.0373]
Green-space ratio	0.2491 [−0.3279, 0.8274]	0.1768 [−0.2445, 0.6765]	0.4258 [−0.5757, 1.4694]
Gas and central-heating availability	0.0571 [−0.0759, 0.1925]	0.0405 [−0.0576, 0.1540]	0.0976 [−0.1327, 0.3404]
Metro distance	−0.0522 [−0.0983, −0.0050]	−0.0371 [−0.0788, −0.0036]	−0.0893 [−0.1716, −0.0088]
Park distance	−0.0093 [−0.0573, 0.0384]	−0.0066 [−0.0437, 0.0301]	−0.0159 [−0.0998, 0.0678]
Shopping-centre distance	−0.0419 [−0.1031, 0.0199]	−0.0297 [−0.0772, 0.0161]	−0.0716 [−0.1750, 0.0362]
Hospital distance	−0.0443 [−0.1116, 0.0224]	−0.0314 [−0.0926, 0.0160]	−0.0757 [−0.2008, 0.0384]
School distance	−0.0187 [−0.0612, 0.0251]	−0.0133 [−0.0471, 0.0192]	−0.0320 [−0.1057, 0.0442]
Opening-year area-level GDP	0.000150 [0.000076, 0.000220]	0.000106 [0.000058, 0.000166]	0.000256 [0.000141, 0.000366]

Notes: $N = 198$. Entries are average spatial effects from the principal SAR model under the row-standardized binary KNN-5 matrix. Simulation-based 95% confidence intervals are reported in brackets and are based on 20,000 joint parameter draws. Distance variables are transformed as $\ln(1 + D)$.

3.3 Multiscale Spatial Heterogeneity

3.3.1 Variable-Specific Spatial Scales

The common-bandwidth GWR model selects an adaptive bandwidth of 136 communities and produces an AICc of 291.004, an R^2 of 0.8254, and an adjusted R^2 of 0.7937. The corresponding MGWR estimates have an AIC of 277.107, an AICc of 287.800, a BIC of 373.900, and an effective number of parameters of 28.436. Its R^2 and adjusted R^2 are 0.8237 and 0.7940, respectively. The adjusted explanatory power of the two models is therefore very similar, while MGWR provides the additional information of a separate bandwidth for each relationship.

Figure 3 shows marked variation in the variable-specific bandwidths. Shopping-centre distance has the smallest bandwidth, at 41 communities, distinguishing it from the other covariates as a highly localized housing-price relationship. Median dwelling floor area has a bandwidth of 122, indicating an intermediate spatial scale. Park distance, green-space ratio, and metro distance form a broader group, with bandwidths

of 145, 147, and 148 communities. Although these relationships draw on information from a substantial portion of the sample, their bandwidths remain below those of the near-global variables.

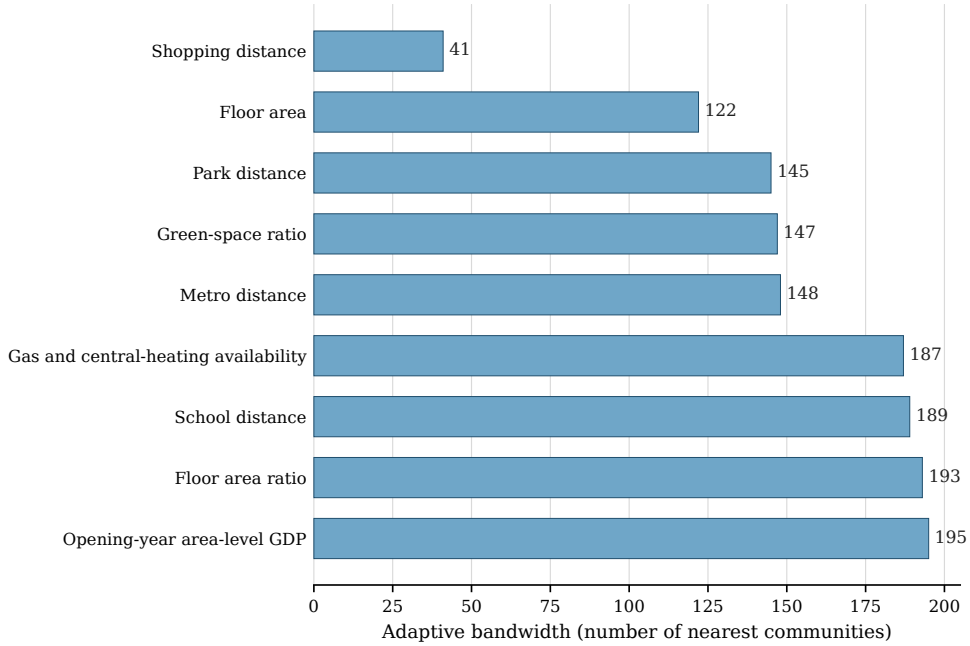


Fig. 3 Variable-specific adaptive bandwidths estimated by MGWR
Notes: Bandwidths indicate the number of nearest communities used to estimate each coefficient surface. Smaller bandwidths represent more localized relationships, while bandwidths approaching the full sample indicate broader spatial scales. The intercept bandwidth is 161 and is omitted from the figure to focus on the explanatory variables.

Gas and central-heating availability and school distance have bandwidths of 187 and 189 communities, while FAR and opening-year area-level GDP have bandwidths of 193 and 195. The GDP relationship is thus estimated at a near-global scale within the sample of 198 communities. A similarly broad bandwidth does not, however, imply a similar degree of statistical support. FAR is not significant at any location after adjustment, whereas GDP is significant at all 198 communities. This contrast provides a particularly clear illustration that an MGWR bandwidth describes the spatial scale of a relationship rather than its statistical significance.

3.3.2 Local Coefficient Patterns

The local coefficient surfaces add information about how the magnitude of each relationship varies within Zhengzhou. Figure 4 maps four variables selected to represent highly local, intermediate, broad, and near-global spatial scales: shopping-centre distance, median dwelling floor area, metro distance, and opening-year area-level GDP. Because the dependent and explanatory variables are standardized before estimation, the mapped coefficients are expressed on comparable standardized scales. Hollow markers identify communities at which the coefficient does not satisfy the variable-specific adjusted 95% significance threshold.

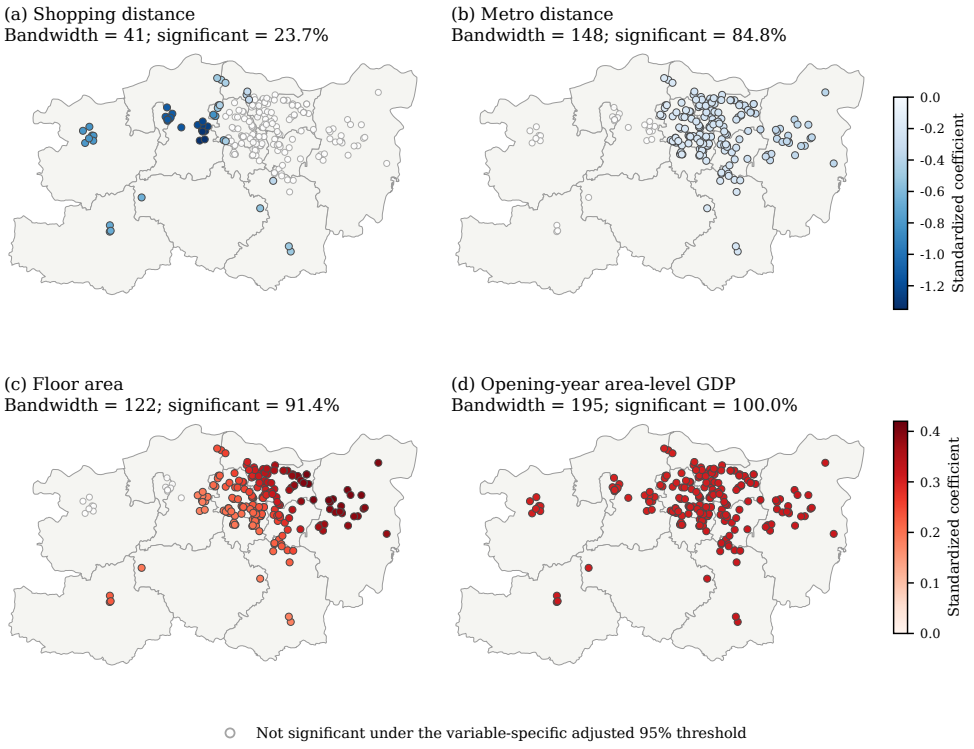


Fig. 4 Spatial variation in selected MGWR coefficients
Notes: The panels show standardized local coefficients for (a) shopping-centre distance, (b) metro distance, (c) median dwelling floor area, and (d) opening-year area-level GDP. Hollow circles indicate coefficients that are not statistically significant under the variable-specific adjusted 95% threshold. Bandwidths are adaptive nearest-neighbour bandwidths.

Shopping-centre distance displays the strongest localization, with standardized local coefficients ranging from -1.347 to 0.229 and averaging -0.282. After adjustment for multiple local tests, 47 communities (23.7%) have statistically supported coefficients, all of which are negative, indicating that the association is concentrated in particular parts of Zhengzhou rather than distributed uniformly across the housing market. Metro distance operates over a broader spatial scale: its coefficients range from -0.387 to 0.065, average -0.225, and are negative and statistically supported at 168 communities (84.8%). Median dwelling floor area occupies an intermediate position in terms of bandwidth and has positive coefficients at all 198 communities, ranging from 0.152 to 0.407 with a mean of 0.273; the adjusted significance criterion is satisfied at 181 communities (91.4%). This consistently positive direction agrees with the dwelling-type baseline and SAR results, while the variation in magnitude shows that the strength of the floor-area relationship is not spatially constant. Opening-year area-level GDP presents a near-global pattern, with a bandwidth of 195 and positive, statistically supported coefficients at every sampled community. The variables not displayed in Figure 4 exhibit other combinations of spatial scale and local support: school distance and gas and central-heating availability are significant at 178 and 158 communities, respectively, whereas green-space ratio and park distance are significant at 78 and 29 communities. FAR has a near-global bandwidth of 193 but is not statistically significant at any location after adjustment, further demonstrating that bandwidth and local statistical significance describe different properties of an estimated relationship.

Local diagnostic results do not indicate severe collinearity in the final MGWR specification. The local condition numbers average approximately 4.95 and reach a maximum of 16.49; 191 communities have values below 15, seven have values between 15 and 30, and none exceed 30. Residual Moran's I is approximately 0.023 and is not statistically significant under the baseline KNN-5 matrix, indicating no detectable residual spatial autocorrelation under this spatial-weights definition.

3.4 Robustness and Sensitivity Results

The spatial results remain stable across alternative spatial weights. Moran's I remains positive and significant as the number of neighbours increases from $K = 3$ to $K = 10$, ranging from 0.7771 to 0.6403. Across the alternative KNN matrices, the robust LM-lag test remains significant, the robust LM-error test does not, and the LR and Wald tests do not reject the SDM-to-SAR restriction. Binary KNN-5 and inverse-distance KNN-5 and KNN-8 produce similar SAR parameters (0.4382, 0.4361, and 0.4593, respectively), no detectable residual spatial autocorrelation, and only limited variation in the principal spatial effects.

The results are also insensitive to the construction of the community-level price measure. Median and mean dwelling-type prices are highly correlated across communities, with Pearson and Spearman correlations above 0.997. Replacing the median with the mean changes the SAR parameter only from 0.4382 to 0.4390, preserves the directions of all substantive coefficients, and leaves no significant residual spatial autocorrelation.

MGWR estimates are similarly stable when the minimum admissible bandwidth is reduced from 40 to 30 communities. The shopping-centre bandwidth changes only from 41 to 40, and the corresponding local coefficients have a correlation of 0.9997. A lower bound of 20 produces a different solution with severe numerical ill-conditioning despite its lower AICc. The similarity of the minimum-40 and minimum-30 results therefore supports the multiscale patterns reported for the principal specification.

4 Discussion

4.1 Substantive Interpretation of the Findings

The Zhengzhou results show that housing-price differences combine spatial dependence with geographically varying relationships. Community prices remain spatially correlated after observed housing, accessibility, and economic characteristics are

controlled for. The SAR specification accounts for the detectable residual spatial autocorrelation under the baseline weights matrix and is consistent with price interdependence or omitted spatially structured influences among neighbouring communities. The cross-sectional model cannot distinguish between these mechanisms. This finding is consistent with spatial hedonic research emphasizing shared locational conditions and spatial connections among housing markets (Anselin and Le Gallo 2006; Cohen and Coughlin 2008; Bhattacharjee et al. 2012).

Floor area and opening-year area-level GDP show positive associations with unit prices in both the dwelling-type baseline and the community-level SAR analysis. Their consistent directions indicate that dwelling size and the economic context at project opening are relevant at both observational scales. The dwelling-type analysis retains variation among the recorded housing configurations, while the community-level analysis identifies relationships among residential developments. Together, they provide complementary evidence on the housing attributes and economic conditions associated with price differences.

Accessibility exhibits a more differentiated pattern. Shopping-centre distance has a negative association in the dwelling-type baseline, while its global SAR effects are not statistically significant. MGWR identifies statistically supported negative local coefficients in 47 communities, indicating that the retail-accessibility relationship is concentrated in particular parts of Zhengzhou. Metro distance shows a more widespread pattern, with negative SAR effects and statistically supported negative MGWR coefficients in 168 communities. The contrast highlights the importance of distinguishing localized retail-accessibility relationships from the broader association between metro proximity and housing prices.

The variable-specific bandwidths further distinguish these spatial patterns. Shopping-centre accessibility is highly localized, the floor-area relationship varies at an intermediate scale, metro accessibility has a broader pattern, and opening-year

area-level GDP shows a near-global positive relationship. These findings are consistent with housing studies documenting spatial variation in accessibility and other price relationships (Liu et al. 2022; Sachdeva et al. 2022; Lu et al. 2023). Although GWR and MGWR have similar adjusted explanatory power, MGWR differentiates the spatial patterns of individual variables. SAR and MGWR therefore contribute complementary evidence: the former characterizes dependence among community prices, while the latter identifies where individual relationships vary and how broadly those patterns extend.

4.2 Implications for Housing Analysis and Urban Policy

The spatial dependence in community prices suggests that coordinated housing-market monitoring across neighbouring areas may be informative. Municipal housing authorities could combine project-level records with community-level price summaries to assess local price differences in their surrounding market context. Consistent community identifiers would preserve information on dwelling composition and provide a common spatial unit for monitoring. Comparisons across adjacent communities, including those on different sides of administrative boundaries, could help distinguish isolated project characteristics from broader geographic price patterns.

The accessibility findings support differentiated planning assessments. The widespread metro relationship suggests that transport access and housing provision should be considered jointly across the urban network. Assessments of station-area development and feeder connections could combine housing distribution with travel demand and existing transport coverage. The localized shopping-centre relationship points to a finer scale of assessment, using neighbourhood-level information on retail provision, resident demand, and accessibility gaps. This distinction provides a basis for matching the geographic scale of planning analysis to the type of accessibility under consideration.

The variation in the floor-area relationship also supports housing-supply assessments that distinguish local submarkets. Authorities could examine dwelling-size composition alongside household structure, purchasing capacity, and local service provision. Accessibility improvements and housing programmes should incorporate affordability and service-equity objectives, with attention to neighbourhoods where residents face limited transport or daily-service access. The spatial results can help organize these assessments and identify locations requiring more detailed investigation.

4.3 Limitations and Future Research

The findings concern platform-displayed filing prices in the sampled communities rather than realized transaction prices. The platform-based, non-probability sample, the cross-sectional design, unobserved housing and neighbourhood characteristics, and dependence on specified spatial weights limit causal interpretation and generalization. In addition, 2026 prices are linked to area-level GDP measured in each project's opening year; this historical matching captures opening-period context but does not identify a contemporaneous income or output effect. Future research could combine transaction or repeated-price data with network-based accessibility measures to examine changes over time. Applications to other cities could further assess the transferability of the two-scale framework.

5 Conclusions

This study develops a two-scale empirical framework for housing-price analysis, centred on the construction of linked dwelling-type and community-level datasets and indicators. At the dwelling-type scale, the framework retains individual prices and housing attributes to estimate baseline conditional relationships. At the community scale, it aggregates dwelling-level information and combines it with shared locational, accessibility, and economic indicators to examine spatial dependence and heterogeneous

relationships. This design aligns indicator construction and model function with the observational scale of each analytical question.

The Zhengzhou application demonstrates the feasibility of this framework. The linked datasets support baseline estimation, spatial econometric analysis, and multiscale modelling within a consistent research design. The findings reveal both dependence among community prices and variation in the spatial patterns of individual relationships, with broadly stable results across the robustness checks examined.

The framework provides a transferable analytical structure for cities with housing records nested within residential communities. Its application requires consistently defined prices, community identifiers, geographic information, and relevant covariates. City-specific model selection and diagnostics allow the explanatory variables, spatial connectivity, and estimated scales to reflect local housing-market conditions. The contribution therefore lies in a reusable two-scale research design, with Zhengzhou providing an empirical demonstration.

References

Anselin L, Le Gallo J (2006) Interpolation of air quality measures in hedonic house price models: Spatial aspects. Spatial Economic Analysis 1(1):31–52. https://doi.org/10.1080/17421770600661337

Anselin L, Lozano-Gracia N (2009) Spatial hedonic models. In: Mills TC, Patterson K (eds) Palgrave Handbook of Econometrics: Volume 2: Applied Econometrics. Palgrave Macmillan, p 1213–1250, https://doi.org/10.1057/9780230244405_26

Bhattacharjee A, Castro E, Marques J (2012) Spatial interactions in hedonic pricing models: The urban housing market of Aveiro, Portugal. Spatial Economic Analysis 7(1):133–167. https://doi.org/10.1080/17421772.2011.647058

Bitter C, Mulligan GF, Dall'erba S (2007) Incorporating spatial variation in housing attribute prices: A comparison of geographically weighted regression and the spatial expansion method. Journal of Geographical Systems 9:7–27. https://doi.org/10.1007/s10109-006-0028-7

Bourassa SC, Cantoni E, Hoesli M (2010) Predicting house prices with spatial dependence: A comparison of alternative methods. Journal of Real Estate Research 32(2):139–160. https://doi.org/10.1080/10835547.2010.12091276

Brunsdon C, Fotheringham AS, Charlton ME (1996) Geographically weighted regression: A method for exploring spatial nonstationarity. Geographical Analysis 28(4):281–298. https://doi.org/10.1111/j.1538-4632.1996.tb00936.x

Chen Y, Yang Q, Geng L, et al (2024) Analysis of factors influencing housing prices in mountain cities based on multiscale geographically weighted regression—demonstrated in the central urban area of Chongqing. Land 13(5):602. https://doi.org/10.3390/land13050602

Cohen JP, Coughlin CC (2008) Spatial hedonic models of airport noise, proximity, and housing prices. Journal of Regional Science 48(5):859–878. https://doi.org/10.1111/j.1467-9787.2008.00569.x

Debrezion G, Pels E, Rietveld P (2007) The impact of railway stations on residential and commercial property value: A meta-analysis. The Journal of Real Estate Finance and Economics 35:161–180. https://doi.org/10.1007/s11146-007-9032-z

Florax RJGM, Folmer H, Rey SJ (2003) Specification searches in spatial econometrics: The relevance of Hendry's methodology. Regional Science and Urban Economics 33(5):557–579. https://doi.org/10.1016/S0166-0462(03)00002-4

Fotheringham AS, Yang W, Kang W (2017) Multiscale geographically weighted regression (mgwr). Annals of the American Association of Geographers 107(6):1247–1265. https://doi.org/10.1080/24694452.2017.1352480

Fotheringham AS, Yu H, Wolf LJ, et al (2022) On the notion of 'bandwidth' in geographically weighted regression models of spatially varying processes. International Journal of Geographical Information Science 36(8):1485–1502. https://doi.org/10.1080/13658816.2022.2034829

Gerkman LM, Ahlgren N (2014) Practical proposals for specifying k-nearest neighbours weights matrices. Spatial Economic Analysis 9(3):260–283. https://doi.org/10.1080/17421772.2014.930167

Gröbel S, Thomschke L (2018) Hedonic pricing and the spatial structure of housing data – an application to Berlin. Journal of Property Research 35(3):185–208. https://doi.org/10.1080/09599916.2018.1510428

Hill RJ (2013) Hedonic price indexes for residential housing: A survey, evaluation and taxonomy. Journal of Economic Surveys 27(5):879–914. https://doi.org/10.1111/j.1467-6419.2012.00731.x

Hu Y, Lu B, Ge Y, et al (2022) Uncovering spatial heterogeneity in real estate prices via combined hierarchical linear model and geographically weighted regression. Environment and Planning B: Urban Analytics and City Science 49(6):1715–1740. https://doi.org/10.1177/23998083211063885

Jin T, Cheng L, Liu Z, et al (2022) Nonlinear public transit accessibility effects on housing prices: Heterogeneity across price segments. Transport Policy 117:48–59. https://doi.org/10.1016/j.tranpol.2022.01.004

Kang W, Oshan TM (2025) Scale and correlation in multiscale geographically weighted regression (MGWR). Journal of Geographical Systems 27(3):399–424. https://doi.org/10.1007/s10109-025-00468-1

Karlsson V (2011) The relationship of housing prices and transportation improvements: Location and marginal impact. Spatial Economic Analysis 6(2):223–241. https://doi.org/10.1080/17421772.2011.557774

Koley M, Bera AK (2024) To use, or not to use the spatial durbin model?—that is the question. Spatial Economic Analysis 19(1):30–56. https://doi.org/10.1080/17421772.2023.2256810

Liu L, Yu H, Zhao J, et al (2022) Multiscale effects of multimodal public facilities accessibility on housing prices based on MGWR: A case study of Wuhan, China. ISPRS International Journal of Geo-Information 11(1):57. https://doi.org/10.3390/ijgi11010057

Liu L, Meng L, Zhang R (2024) Does easy accessibility to urban parks always raise home values? Journal of Housing Economics 63:101981. https://doi.org/10.1016/j.jhe.2023.101981

Lu B, Ge Y, Shi Y, et al (2023) Uncovering drivers of community-level house price dynamics through multiscale geographically weighted regression: A case study of Wuhan, China. Spatial Statistics 53:100723. https://doi.org/10.1016/j.spasta.2022.100723

Oshan TM, Li Z, Kang W, et al (2019) mgwr: A Python implementation of multiscale geographically weighted regression for investigating process spatial heterogeneity and scale. ISPRS International Journal of Geo-Information 8(6):269. https://doi.org/10.3390/ijgi8060269

Osland L (2010) An application of spatial econometrics in relation to hedonic house price modeling. Journal of Real Estate Research 32(3):289–320. https://doi.org/10.1080/10835547.2010.12091282

Rosen S (1974) Hedonic prices and implicit markets: Product differentiation in pure competition. Journal of Political Economy 82(1):34–55. https://doi.org/10.1086/260169

Sachdeva M, Fotheringham S, Li Z (2022) Do places have value?: Quantifying the intrinsic value of housing neighborhoods using MGWR. Journal of Housing Research 31(1):24–52. https://doi.org/10.1080/10527001.2021.2003505

Wen H, Xiao Y, Hui ECM, et al (2018) Education quality, accessibility, and housing price: Does spatial heterogeneity exist in education capitalization? Habitat International 78:68–82. https://doi.org/10.1016/j.habitatint.2018.05.012

Yang L, Chau KW, Szeto WY, et al (2020) Accessibility to transit, by transit, and property prices: Spatially varying relationships. Transportation Research Part D: Transport and Environment 85:102387. https://doi.org/10.1016/j.trd.2020.102387

Yang L, Liang Y, He B, et al (2022) COVID-19 effects on property markets: The pandemic decreases the implicit price of metro accessibility. Tunnelling and Underground Space Technology 125:104528. https://doi.org/10.1016/j.tust.2022.104528

Zhou Y, Tian Y, Jim CY, et al (2022) Effects of public transport accessibility and property attributes on housing prices in polycentric Beijing. Sustainability 14(22):14743. https://doi.org/10.3390/su142214743

Data Availability

The housing dataset used in this study, the processed analytical datasets, the Python analysis code, and supporting documentation are provided in the accompanying supplementary archive, "data and supporting documentation.zip". These materials are supplied to support verification and reproduction of the analyses by the editors and reviewers.

Ethical Approval

Not applicable. This article does not contain any studies with human participants performed by any of the authors.

Informed Consent

Not applicable. Informed consent was not applicable because this study did not involve human participants.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AuthorIdentityandAffiliationDocumentation.pdf
- dataandsupportingdocumentation.zip