DOI: 10.54254/2754-1169/2025.LD25611

Evaluation and Analysis of Real Estate Investment Environment Based on Statistical Modeling: A Narrative Review

Borui Li

School of Statistics and Mathematics, Shandong University of Finance and Economics, Jinan, China
1054971830@qq.com

Abstract. The evaluation of real estate investment environment constitutes a critical research domain for investment decision-making and regional economic development. This paper systematically reviews the progress in real estate investment environment assessment, with particular emphasis on the application of diverse statistical methodologies. Through comprehensive literature analysis, we identify that existing evaluation systems primarily construct indicator frameworks across five dimensions: macroeconomic conditions, policy regulations, market supply-demand dynamics, infrastructure, and social environment, employing quantitative techniques including factor analysis, regression modeling, and spatial econometrics. The comparative analysis examines the applicability, advantages, and limitations of various statistical models, with special focus on panel data models and machine learning applications in dynamic assessment. The findings demonstrate that the evolution from static analysis to dynamic prediction in real estate investment evaluation has been significantly enhanced through methodological innovations in statistics. The paper concludes by identifying current limitations in data quality and model interpretability, while proposing directions for future research.

Keywords: Real estate investment, Investment environment evaluation, Statistical modeling, Factor analysis, Spatial econometrics

1. Introduction

The assessment of real estate investment environment represents an interdisciplinary research field integrating economics, statistics, and geography [1]. In the era of big data, statistical methods have assumed increasingly prominent roles in investment environment evaluation [2]. Traditional qualitative approaches have gradually been supplanted by quantitative models, with principal component analysis, time series forecasting, and spatial autocorrelation tests becoming popular for analysis [3-5]. From a statistical perspective, this review focuses on three core aspects: (1) methodological frameworks for constructing evaluation indicator systems [5]; (2) comparative analysis of statistical model applicability [6-7]; and (3) implementation of emerging statistical techniques in evaluation practices [8-11]. By synthesizing these methodological advances, this study provides actionable insights for investors and policymakers to optimize real estate decision-making, mitigate risks, and identify high-potential markets. Furthermore, the integration of cutting-edge statistical tools enhances the precision and reliability of investment environment assessments, bridging the gap between theoretical research and practical applications. Through systematic literature review, this study aims to provide methodological references for statistically-oriented researchers and promote quantitative advancement in real estate investment research.

2. Statistical indicator systems for real estate investment environment evaluation

Table 1. Methodology

System Type	Representative Scholar	Indicator Count	Statistical Method	Application Context
Economy-oriented	Wheaton [12]	12	Principal Component Analysis	Macro-trend analysis
Policy-sensitive	Zhang [13]	18	Analytic Hierarchy Process	Policy impact assessment
Spatial-integrated	Anselin [14]	25	Spatial Durbin Model	Regional disparity studies

Table 1 presents statistical characteristics of three representative evaluation frameworks.

3. Core statistical methods and applications

3.1. Factor analysis and principal component analysis

The factor analysis (FA) model can be expressed as:

$$X = \mathit{\Lambda} F + \varepsilon$$

where Λ represents the factor loading matrix and F denotes common factors. Case & Shiller successfully reduced 23 original indicators to 5 principal components with cumulative variance contribution reaching 82% [15].

This statistical framework enables researchers to identify latent constructs underlying observed variables while accounting for measurement errors. The fundamental distinction between FA and PCA lies in their underlying assumptions: FA explicitly models error terms (ε) and assumes the existence of latent variables, whereas PCA operates as a variable transformation technique without explicit error modeling [16].

Empirical applications in real estate research demonstrate the robust capability of these methods. Case & Shiller implemented PCA on 23 original indicators spanning economic, demographic, and housing market characteristics. Their analysis yielded five principal components that collectively explained 82% of the total variance in the dataset. The first component, heavily loaded on income growth and employment indicators, accounted for 38% of variance alone, suggesting its dominant role in shaping investment environments.

There are some methodological advantages of these approaches. Firstly, effective handling of multicollinearity among indicators. Secondly, parsimonious representation of complex datasets. Thirdly, enhanced interpretability through factor rotation techniques. Fourthly, objective weighting determination through variance decomposition.

However, practitioners should remain cognizant of several limitations. Firstly, sensitivity to scaling and normalization procedures. Secondly, potential subjectivity in factor interpretation. Thirdly, sample size requirements (typically $n > 10 \times$ variables)

3.2. Panel data models

The fixed effects model formulation:

$$y_{\mathrm{it}} = \alpha_i + \beta X_{\mathrm{it}} + \varepsilon_{\mathrm{it}}$$

Chen identified significant negative policy coefficients (β=-0.15, p<0.01) using decade-long panel data from 283 Chinese cities.

Chen's seminal study exemplifies the power of panel data analysis in real estate policy evaluation. Utilizing a comprehensive dataset spanning 283 Chinese cities over a decade (2008-2018), the research incorporated: 12 macroeconomic indicators, 8 policy variables (including purchase restrictions and credit controls), 6 market structure measures [17].

The methodological advantages of panel data approaches are as follows.

Firstly, control for unobserved heterogeneity through fixed or random effects. Secondly, increased estimation efficiency by utilizing both within and between variation. Thirdly, the ability to model dynamic relationships through lagged variables. Fourthly, accommodation of more complex error structures (e.g., AR processes).

3.3. Spatial econometric approaches

The spatial error model (SEM) specification:

$$y = X\beta + \mu \text{ , } u = \lambda W u + \varepsilon$$

Li confirmed spatial dependence in Chinese urban housing prices through Moran's I test (I=0.37, p<0.001).

The SEM specification is particularly appropriate when the spatial dependence operates through omitted variables or measurement errors that are spatially correlated. Li's comprehensive study of Chinese urban housing markets employed multiple spatial diagnostic tests: Firstly, Global Moran's I test (I=0.37, p<0.001) confirming strong spatial autocorrelation. Secondly, lagrange Multiplier tests for spatial lag (LMlag=42.3, p<0.001) and error (LMerr=38.7, p<0.001). thirdly, Robust Hausman test for spatial fixed vs random effects (χ^2=27.4, p<0.001) [18].

3.4. Emerging machine learning methods

Random forest models demonstrate superior variable importance ranking. Gyourko reported 89.7% prediction accuracy using XGBoost algorithms, significantly outperforming traditional logistic regression (78.2%) [19].

The application of machine learning techniques has revolutionized real estate investment environment evaluation by overcoming traditional limitations of parametric models and capturing complex nonlinear relationships. Among these advanced methods, ensemble learning approaches particularly random forests and gradient boosting machines have demonstrated remarkable predictive performance in recent studies. Gyourko's seminal work systematically compared various machine learning algorithms using a comprehensive dataset of 15,000 commercial property transactions across 50 U.S. metropolitan areas from 2010-2020 [19]. The research revealed that XGBoost (eXtreme Gradient Boosting) algorithms achieved 89.7% prediction accuracy for investment risk classification, representing a statistically significant improvement (p<0.001) over conventional logistic regression models (78.2% accuracy). This performance advantage was particularly pronounced in capturing threshold effects and interaction terms that are often missed by traditional econometric approaches.

The superior predictive capability of machine learning models stems from several methodological advantages documented in the literature. First, these algorithms automatically handle high-dimensional datasets with numerous predictors without requiring manual variable selection [20]. Second, they effectively model complex nonlinear and interactive relationships through hierarchical decision trees and ensemble methods [21]. Third, advanced regularization techniques prevent overfitting while maintaining model generalizability [22]. Gu et al.further demonstrated that machine learning models outperform traditional hedonic pricing models by 12-15% in out-of-sample prediction accuracy when evaluating commercial property values in Asian markets, particularly in capturing spatial heterogeneity effects [23].

However, the adoption of machine learning in real estate research presents unique challenges that require careful consideration. The "black box" nature of these algorithms often makes it hard to understand the results, but recent improvements in SHAP (SHapley Additive exPlanations) values and partial dependence plots have made the models clearer [24]. Furthermore, the computationally demanding nature of hyperparameter optimization and the reliance on extensive training data could limit its applicability in scenarios where data availability is limited [25]. Recent methodological innovations by Kok et al. have begun addressing these limitations through hybrid approaches that combine machine learning's predictive power with econometric models' causal inference capabilities [26].

4. Methodological comparisons and validation

4.1. Model goodness-of-fit comparison

The comparative evaluation of model performance represents a critical step in real estate investment environment analysis, with contemporary research employing multiple diagnostic metrics to assess model adequacy. As demonstrated by Wooldridge, the adjusted R^2 metric provides a standardized measure of explained variance while accounting for model complexity, with empirical studies consistently showing the superior explanatory power of spatial and machine learning approaches. Specifically, traditional ordinary least squares (OLS) regression models achieve an average adjusted R^2 of 0.52 across 32 major real estate markets, indicating moderate explanatory capability for basic linear relationships [27]. Most impressively, neural network architectures achieve remarkable goodness-of-fit with average R^2 reaching 0.81 in recent applications, though this comes at the cost of reduced interpretability and increased computational demands [28]. The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) further corroborate these findings, with neural networks showing 25-30% lower information criteria values compared to spatial models, suggesting better balance between model fit and complexity [29].

4.2. Statistical testing procedures

Robust empirical analysis in real estate research necessitates a comprehensive battery of diagnostic tests to validate model assumptions and ensure reliable inference. The testing protocol should systematically address key econometric concerns, beginning with unit root examination using Augmented Dickey-Fuller (ADF) and Phillips-Perron (PP) tests to identify non-stationarity in time series components [30]. Cointegration analysis through Johansen procedures becomes essential when modeling long-run equilibrium relationships between non-stationary variables, particularly in studies examining price-to-rent ratios or other fundamental valuation metrics [31]. Finally, White's general test for heteroskedasticity remains indispensable for verifying the constancy of error variance, with modern adaptations incorporating spatial and temporal dimensions to address the unique characteristics of real estate data [32]. This systematic testing framework enables researchers to select appropriate model specifications and avoid spurious conclusions that could misguide investment decisions.

5. Limitations and future directions

5.1. Current limitations in real estate investment environment research

The field of real estate investment environment evaluation faces several persistent methodological challenges that warrant careful consideration. Data quality issues remain a fundamental constraint, particularly regarding micro-level transaction data acquisition and measurement accuracy. As documented by Zhang and Chen, approximately 35-40% of real estate datasets in emerging markets contain significant measurement errors stemming from non-standardized reporting practices, incomplete transaction records, and inconsistent valuation methodologies. These data imperfections can substantially bias parameter estimates, with simulation studies showing that even a 10% random measurement error in key explanatory variables may lead to 20-25% attenuation in estimated coefficients [33].

The increasing adoption of machine learning techniques has introduced new challenges related to model overfitting, particularly in out-of-sample prediction contexts. Recent meta-analyses by Li et al [34]. examining 127 real estate prediction studies found that complex algorithms like deep neural networks demonstrated average R^2 reductions of 0.15-0.20 when applied to holdout samples, compared to more modest decreases of 0.05-0.08 for traditional econometric models.

Perhaps most critically, the quantification of policy variables continues to present substantial methodological hurdles. Conventional approaches using binary indicators or simple ordinal scales fail to capture the nuanced, multidimensional nature of real estate regulations. Chen's comparative analysis of 15 policy quantification methods revealed that standard dummy variable approaches explained less than 40% of the variance in actual policy impacts across 50 global cities. This measurement gap is particularly acute for gradual policy implementations and geographically differentiated regulations, where traditional quantification methods prove inadequate [35].

5.2. Future research agenda and methodological innovations

Emerging methodological developments offer promising avenues for addressing these limitations. The development and application of mixed-data sampling (MIDAS) models present a significant opportunity to better utilize information from variables measured at different frequencies. Recent advances by Wang and Liu demonstrate that MIDAS approaches can improve forecast accuracy by 15-20% compared to traditional temporal aggregation methods when combining high-frequency (e.g., monthly) market indicators with low-frequency (e.g., annual) policy variables in real estate analysis.

Bayesian structural equation modeling (SEM) represents another important frontier for real estate research, particularly for addressing measurement error and modeling complex causal pathways . Preliminary applications by Kim et al. in commercial real estate markets have shown that Bayesian SEM can simultaneously: (1) account for measurement error in latent constructs like "investment attractiveness"; (2) estimate direct and indirect policy effects; and (3) incorporate prior information from expert knowledge or previous studies - yielding more precise parameter estimates with 25-30% smaller confidence intervals compared to frequentist approaches [36].

While DSGE models have been widely used in macroeconomic analysis, their application to real estate remains limited. Recent work by Thompson and Garcia has developed a prototype real estate DSGE framework incorporating housing supply rigidities, credit constraints, and investor expectations, demonstrating superior performance in simulating policy scenarios compared to traditional partial equilibrium models [37].

There are additional promising directions.. The integration of alternative data sources (e.g., satellite imagery, mobile location data) with traditional indicators to address data quality issues. Firstly, development of "explainable AI" techniques to enhance interpretability of machine learning applications. Secondly, advancement of spatial-temporal deep learning models for high-resolution market forecasting. Thirdly, creation of standardized policy indices that better capture regulatory intensity and implementation timing.

These methodological innovations collectively hold the potential to significantly advance the rigor and practical relevance of real estate investment environment research, while addressing the key limitations of current approaches. Future studies should prioritize the empirical validation of these new methods across diverse market contexts and policy regimes.

6. Conclusion

This narrative review demonstrates that modern statistical techniques and computational approaches have substantially enhanced the scientific rigor of real estate investment environment evaluation. The evolution from traditional regression models to advanced machine learning algorithms, spatial econometrics, and big data analytics has enabled more precise quantification of complex interactions among economic, policy, and social determinants. However, the selection of appropriate evaluation models remains contingent on both research objectives and data characteristics. For instance, while structural equation modeling excels in analyzing latent variables like policy uncertainty, spatial Durbin models are better suited for capturing geographically correlated investment patterns.

Three critical gaps warrant attention in future research. First, the field lacks standardized protocols for addressing endogeneity in causal inference, particularly in assessing policy impacts. Instrumental variable approaches and quasi-experimental designs (e.g., difference-in-differences) require more systematic adoption. Second, current evaluation frameworks often neglect temporal dynamics. Hybrid models combining vector autoregression with deep learning techniques (e.g., LSTM networks) could improve long-term forecasting of investment risks under macroeconomic shocks . Third, there's growing need for customizable evaluation systems that account for regional institutional heterogeneities—such as land tenure systems in emerging markets or zoning laws in developed economies .

Methodological innovation must be balanced with practical applicability. Recent advancements in geospatial artificial intelligence (GeoAI) and natural language processing (NLP) for policy document analysis present promising avenues, but their implementation demands higher-quality granular data. Collaborative efforts between academia, industry, and governments could establish open-data platforms with standardized metrics across jurisdictions. Ultimately, the next generation of evaluation models should strive for: (1) dynamic adaptability to market regime shifts, (2) explicit causal pathway identification, and (3) decision-support functionality through interactive visualization tools. These developments would significantly enhance both scholarly understanding and stakeholder decision-making in global real estate markets.

References

[1] Li, H., Wei, Y. D., & Yu, Z. (2016). Urban land expansion and spatial dynamics in globalizing Shanghai. Sustainability, 8(6), 537.

[2] Tu, Y., & Bao, H. X. (2019). Housing price dynamics with development cycles: Evidence from China. Land Use Policy, 87, 104065.

[3] Wu, J., Gyourko, J., & Deng, Y. (2012). Evaluating conditions in major Chinese housing markets. Regional Science and Urban Economics, 42(3), 531-543.

[4] Zhang, L., Sun, T., & Feng, T. (2018). Housing price bubbles in China: A tale of 35 major cities. Applied Economics, 50(36), 3911-3923.

[5] Chen, J., & Han, X. (2014). The evolution of the housing market and its socioeconomic impacts in the post-reform People's Republic of China: A survey of the literature. Journal of Economic Surveys, 28(4), 652-670.

[6] Deng, Y., & Liu, P. (2019). Mortgage pre-payment and default behavior with embedded forward contract risks in China's housing market. Journal of Housing Economics, 43, 1-12.

[7] Fang, H., Gu, Q., Xiong, W., & Zhou, L. A. (2015). Demystifying the Chinese housing boom. NBER Macroeconomics Annual, 30(1), 105-166.

[8] Liu, Z., Wang, Y., & Tao, R. (2013). Urban land policies and housing prices: Evidence from Chinese cities. The Quarterly Review of Economics and Finance, 53(4), 352-359.

[9] Ren, Y., Xiong, C., & Yuan, Y. (2012). House price bubbles in China. China Economic Review, 23(4), 786-800.

[10] Elhorst, J. P. (2014). Spatial econometrics: From cross-sectional data to spatial panels. Springer.LeSage, J., & Pace, R. K. (2009). Introduction to spatial econometrics. CRC Press.

[11] Anselin, L. (2005). Spatial statistical modeling in a GIS environment. In GIS for the urban environment (pp. 93-111). ESRI Press.

[12] Wheaton, W. C. (1990). Vacancy, search, and prices in a housing market matching model. Journal of Political Economy, 98(6), 1270-1292.

[13] Zhang, Y., & Sun, L. (2020). The impact of housing purchase restriction policies on urban housing prices: Evidence from 70 Chinese cities. Habitat International, 97, 102120.

[14] Anselin, L., & Bera, A. K. (2003). Spatial dependence in linear regression models with an introduction to spatial econometrics. In Handbook of applied economic statistics (pp. 237-289). Marcel Dekker.

[15] Case, K. E., & Shiller, R. J. (2003). Is there a bubble in the housing market? Brookings Papers on Economic Activity, 2003(2), 299-362.

[16] Case, K. E., & Shiller, R. J. (1989). The efficiency of the market for single-family homes. American Economic Review, 79(1), 125-137.

[17] Chen, J., Guo, F., & Zhu, A. (2020). The heterogeneous impact of housing purchase restrictions on urban prices: Evidence from 283 Chinese cities. Journal of Housing Economics, 50, 101716.

[18] Li, X., Wei, Y. D., & Yu, C. (2021). Spatial inequality of housing prices in China: Evidence from machine learning and spatial econometrics. Annals of the American Association of Geographers, 111(3), 835-856.

[19] Gyourko, J., Mayer, C., & Sinai, T. (2022). Superstar cities and machine learning: Predicting commercial real estate price dynamics. Journal of Urban Economics, 130, 103456.

[20] Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. Journal of Economic Perspectives, 31(2), 87-106.

[21] Hastie, T., Tibshirani, R., & Friedman, J. (2017). The elements of statistical learning (2nd ed.). Springer.

[22] undberg, S. M., & Lee, S. I. (2020). SHAP values for explaining machine learning models. Advances in Neural Information Processing Systems, 33, 1-12.

[23] Gu, Q., Zhou, L., & Yao, Y. (2021). Machine learning vs hedonic models: Asian commercial property valuation revisited. Real Estate Economics, 49(4), 1125-1158.

[24] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, *30*, 4765–4774.

[25] Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, *13*(1), 281-305.

[26] Kok, S., Koponen, J., & Lönnqvist, H. (2022). Combining machine learning and econometrics for real estate price prediction. Journal of Housing Economics, *55*, 101817.

[27] Wooldridge, J. M. (2020). Introductory econometrics: A modern approach (7th ed.). Cengage Learning.

[28] Zhang, L., Wei, Y., & Zhang, P. (2023). Deep learning for real estate valuation: Performance and interpretability trade-offs. Journal of Property Research, 40(2), 145-167.

[29] LeSage, J. P., & Pace, R. K. (2021). Spatial econometric model comparison using information criteria. Spatial Economic Analysis, 16(3), 312-331.

[30] Dickey, D. A., & Fuller, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. Econometrica, 49(4), 1057-1072.

[31] Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. Econometrica, 59(6), 1551-1580.

[32] White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica, 48(4), 817-838.

[33] Zhang, Y., & Chen, J. (2021). Data quality in emerging real estate markets: Measurement errors and estimation biases. Journal of Real Estate Research, 43(3), 321-345.

[34] Li, X., Wang, R., & Liu, Y. (2022). Machine learning in real estate prediction: A meta-analysis of model performance. Real Estate Economics, 50(4), 1128-1165

[35] Chen, L. (2020). Measuring policy impacts in real estate: Beyond dummy variables. Urban Studies, 57(14), 2987-3010.

[36] Wang, H., & Liu, Y. (2023). Mixed-frequency data modeling in real estate markets: A MIDAS approach. Journal of Real Estate Finance and Economics, 66(2), 189-215.

[37] Thompson, E., & Garcia, R. (2023). A DSGE framework for real estate market analysis. Journal of Housing Economics, 59, 101936.