Assessing Residential Land Efficiency with Spatial–Contextual GMM and Human Activity Big Data: A Case Study of Shenzhen

Shihao Liang[1], Yixin Liu[1], Renzhong Guo[1], Weixi Wang[1], Ding Ma[1], Ye Zheng[2], Shengjun Tang[1],Linfu Xie[1],Xiaoming Li [1,*]

1 Research Institute for Smart Cities & MNR Key Laboratory of Urban Land Resources Monitoring and Simulation, School of Architecture and Urban Planning , Shenzhen University, Shenzhen 518060, China
[2] Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo, 315211, China

Keywords: Inefficient urban land, Mobile signaling data, Multi-source spatial data, unsupervised clustering, Urban redevelopment.

Abstract

As China’s urban development shifts toward stock-based optimisation, identifying inefficient residential land has become important for urban regeneration. Existing approaches often rely on subjective weighting, linear analytical structures, or homogeneous treatment of different residential types, which weakens robustness and transferability. This study proposes a data-driven framework that integrates mobile-phone signaling and other multi-source spatiotemporal big data in Shenzhen. Two dominant residential forms—formal residential communities and urban villages—are evaluated separately through a four-dimensional framework covering built form, activity vitality, economic efficiency, and environmental livability. Principal component analysis is used to estimate intrinsic dimensionality and initialize a parametric autoencoder. A spatially constrained Gaussian mixture model is then employed to identify inefficient residential clusters while preserving local coherence. The clustering results are interpreted using a random forest model and TreeSHAP, and externally validated by street-view imagery interpretation and limited field surveys. PCA retained five components for urban villages and six for formal residential communities, and the BIC selected six and five clusters for the two residential types, respectively. The results indicate that inefficient formal residential communities show scattered and island-like spatial patterns, whereas inefficient urban villages tend to form more continuous clusters along the edges of larger village agglomerations. Random forest and TreeSHAP further reveal that inefficient urban villages are more strongly associated with deficiencies in service accessibility and local socioeconomic conditions, whereas inefficient formal residential communities are more closely associated with lower residential vitality and relatively high development intensity. External validation indicates acceptable agreement with observed residential conditions.

1. Introduction

China’s urban development is undergoing a profound transformation from incremental land expansion to stock-based optimisation. Under the pressure of land scarcity, population growth, and rising expectations for living quality, urban regeneration has become a central strategy for improving the efficiency and quality of built-up areas. In this context, the identification of inefficient urban land has become a prerequisite for evidence-based regeneration and spatial governance. This broader transition is consistent with recent discussions on land-use transition, land-use efficiency, and stock-oriented governance in both China and the international literature (Long, 2014; Chen et al., 2020; Estoque et al., 2021; Schiavina et al., 2022; Liu et al., 2020; Wang et al., 2025). Compared with industrial and commercial land, inefficient residential land is more closely related to residents’ daily life, social welfare, and housing quality. Its accurate identification is therefore not only a land-use issue, but also a matter of improving urban livability and supporting equitable renewal.

In this study, residential land efficiency does not refer to a single density- or output-oriented measure. Instead, it is defined as the degree to which a residential parcel achieves a relatively balanced and effective combination of built form, actual residential vitality, economic performance, and surrounding service support. Accordingly, inefficient residential land refers to parcels that show systematic underperformance or imbalance across these dimensions rather than merely exhibiting an extreme value in one indicator. This understanding is also consistent with broader discussions of land resource-use efficiency that emphasize the relationship between land-use performance and the effective use of spatial and ecological resources (Herzig et al., 2018).

A growing body of research has explored the identification of inefficient urban land from the perspectives of land output, spatial morphology, facility provision, and socioeconomic performance. In the case of residential land, scholars have examined dimensions such as building quality, infrastructure conditions, accessibility to public services, environmental quality, and economic value. Related studies on inefficient urban land and inefficient residential land have shown that land-use differentiation and residential heterogeneity should be explicitly considered in regeneration-oriented identification frameworks (Qiao and Huang, 2023; Jin et al., 2023; Wang et al., 2024; Qi et al., 2025). At the same time, recent studies on large-scale urban spatial identification suggest that multi-source geospatial data and automated recognition frameworks can substantially improve the objectivity and scalability of urban space diagnosis (Du et al., 2024; Sun et al., 2023). However, three limitations remain. First, many existing approaches rely on expert scoring or subjective weighting schemes, which may reduce the objectivity and transferability of evaluation results. Second, some studies mainly employ linear analytical structures, making it difficult to capture nonlinear interactions among multidimensional urban features. Third, residential land is often treated as a homogeneous category, although formal residential communities and urban villages differ substantially in built form, occupancy patterns, socioeconomic characteristics, and service environments.

Recent advances in geospatial big data and machine learning provide new opportunities for addressing these limitations.

Mobile-phone signaling data can reveal dynamic patterns of human activity and nighttime occupancy, while multi-source spatial data have been increasingly used to characterize urban spatial structure and functional differentiation (Du et al., 2024; Dong et al., 2023). Building census data, housing-market information, and POI/AOI data can further reflect development intensity, economic performance, and environmental support conditions. At the methodological level, unsupervised learning allows latent patterns of land-use efficiency to be detected without imposing strong prior assumptions, while interpretable machine learning can help explain the factors associated with identified inefficiency.

Against this background, this study develops a data-driven framework to identify inefficient residential land in Shenzhen, China. Formal residential communities and urban villages are analyzed separately to better accommodate residential heterogeneity. Based on multi-source data, a four-dimensional indicator system is constructed to characterize built form, activity vitality, economic efficiency, and environmental livability. After preprocessing and dimensionality reduction, a spatially constrained Gaussian mixture model is used to identify latent residential clusters, from which inefficient types are determined. The main objectives of this study are threefold: (1) to construct a multidimensional and data-driven framework for evaluating residential land efficiency; (2) to identify inefficient residential land through spatially informed soft clustering; and (3) to reveal the spatial patterns and associated factors of inefficient residential land for different residential types. The study is expected to provide methodological support for differentiated urban regeneration and more refined land-use governance.

The main contributions of this study are as follows. First, it explicitly incorporates residential heterogeneity by analyzing formal residential communities and urban villages in parallel rather than under a single framework. Second, it combines PCA-guided representation learning with spatially constrained Gaussian mixture modeling, enabling the identification of inefficient residential land while retaining both nonlinear feature structure and spatial coherence. Third, it links clustering results with post hoc interpretation and external validation, thereby improving the explainability and practical credibility of the identified inefficient land patterns.

2. Data and Indicators

2.1 Study area

Shenzhen is located on the eastern side of the Pearl River Delta and is one of China's most representative megacities in the transition from rapid expansion to stock-based redevelopment. Over the past several decades, the city has experienced sustained economic growth, massive in-migration, and intensive construction, leading to high land development intensity and increasingly scarce developable land resources. Under these conditions, urban regeneration has become a key strategy for improving land-use efficiency and living quality.

At the same time, Shenzhen exhibits strong residential heterogeneity. Alongside formal residential communities developed through standardized planning and market-oriented housing production, a large number of urban villages continue to provide affordable housing for diverse population groups. These two residential forms differ significantly in their building morphology, density, facility provision, population structure, and market performance. Such heterogeneity makes Shenzhen a suitable case for examining the differentiated identification of inefficient residential land.

This study conducts a citywide analysis across Shenzhen using formal residential communities and urban villages as the basic analytical units. This full-city perspective allows inefficient residential land to be examined under the broader metropolitan spatial structure rather than within a single district.

2.2 Data sources

This study integrates multi-source geospatial and socioeconomic data to evaluate residential land efficiency. Formal residential-community parcels and urban villages are used as the basic spatial units of analysis. The main datasets, summarized in Table 1, include:

1. Residential parcel boundaries. Official or open-data vector boundaries of formal residential communities and urban villages are used to define the analytical units.
2. Building census data. Building inventory data provide information such as gross floor area and building coverage characteristics, which are used to derive indicators of built form and development intensity.
3. Housing listing data and rental reference price data are used to derive building-age information and rental performance, respectively, thereby representing the physical and economic dimensions of residential land efficiency.
4. POI and AOI data. Points and areas of interest are used to characterize the availability of transport, medical, educational, and leisure services around each residential parcel.
5. Mobile-phone signaling data. Anonymized mobile-phone signaling records from 12 and 13 June 2024 are used to calculate average nighttime population density. Population-tag attributes from the same data source are further used to derive the median resident income adopted in the final model.
6. street-view imagery and field observations. These data are used for external validation of the identified inefficient residential land.

Dataset	Data Source	Year
Building census	Shenzhen Housing and Construction Bureau	2019
Housing listing data	Anjuke Platform (https://shenzhen.anjuke.com/)	2025
POI	Amap (www.amap.com)	2022
AOI		2024
Mobile-phone signaling data	A major domestic location-data service provider	2024
Population-tag data		
Rental reference price	Shenzhen Government Data Open Platform	2023
Administrative boundary	Shenzhen Planning and Natural Resources Bureau	2024
Street view image	Baidu Map Open Platform	2024

Table 1. Description of datasets

2.3 Indicator system

To capture the multidimensional nature of residential land efficiency, this study first establishes a candidate indicator

https://doi.org/10.5194/isprs-archives-XLIX-B4-2026-167-2026

system from four dimensions: built form, activity vitality, economic efficiency, and environmental livability.

Built form reflects the intensity and physical organisation of land development. The candidate indicators include floor area ratio (FAR), building density (BD), and building age (AGE).

Activity vitality reflects the occupancy and human-use intensity of residential space. Based on mobile-phone signaling records on 12 and 13 June 2024, average nighttime population density (NPD) and the edge-core nighttime population ratio (ECI_NPD) are used to represent actual residential vitality and relative neighborhood vitality.

Economic efficiency reflects the market and socioeconomic performance of residential land. Rental price (RENT), median resident income (MRI), and median resident consumption (MRC) are initially included in this dimension, with the latter two derived from population-tag data.

Environmental livability reflects the quality of surrounding service provision and residential support conditions.

Accessibility-related indicators include medical service accessibility (MSA), educational service accessibility (ESA), leisure and entertainment accessibility (LEA), and public transport accessibility (PTA).

Table 2 presents the initial candidate indicator system, including each indicator’s dimension, description, and calculation method. After correlation screening, 11 indicators are retained for the final citywide experiment. The only excluded variable is MRC, which is removed because of its strong redundancy with MRI.

Dimension	Indicator	Description	Calculation method	Final use
Built form	FAR	Floor area ratio	Total building floor area within parcel / parcel area	Retained
	BD	Building density	Total building footprint area within parcel / parcel area	Retained
	AGE	Building age	Building completion year extracted for each parcel/building and aggregated to parcel level	Retained
Activity vitality	NPD	Average nighttime population density	Average nighttime resident population derived from mobile-phone signaling records on 12 and 13 June 2024 / parcel area	Retained
	ECI_NPD	Edge-core nighttime population ratio	Parcel NPD divided by the median NPD of surrounding parcels	Retained
Economic efficiency	RENT	Rental-price indicator	Mean or median unit-area rent calculated from parcel-level or nearby rental samples	Retained
	MRI	Median resident income	Median income of resident population within each parcel derived from population-tag data	Retained
	MRC	Median resident consumption	Median consumption of resident population within each parcel derived from population-tag data	Removed after correlation screening
Environmental livability	MSA	Medical service accessibility	Number of medical POIs within a 500 m buffer around each parcel	Retained
	ESA	Educational service accessibility	Number of educational POIs within a 500 m buffer around each parcel	Retained
	LEA	Leisure and entertainment accessibility	Number of leisure and entertainment POIs within a 500 m buffer around each parcel	Retained
	PTA	Public transport accessibility	Number of transport stations, including bus and metro stops, within a 500 m buffer around each parcel	Retained

Table 2. Candidate indicator system, calculation methods, and final retention

3. Methodology

As illustrated in Figure 1, the methodological framework consists of five major stages: data preprocessing and feature screening, representation learning, spatial clustering, inefficient-cluster identification, and interpretation and validation.

3.1 Data preprocessing and feature screening

Before model construction, all indicators are cleaned and standardized to improve comparability and statistical stability. Because the indicators are derived from heterogeneous sources and measured in different units, a standardization procedure is first applied. For variables with strong skewness, appropriate transformations, such as logarithmic, square-root, or Yeo–Johnson transformation, are performed to reduce distributional bias and improve the suitability of the data for subsequent Gaussian-based clustering.

To reduce redundancy among indicators, pairwise relationships are examined using the Spearman rank correlation coefficient. When two indicators show strong collinearity, one of them is removed according to interpretability and practical relevance. In the final citywide experiment, a threshold of |rho|>0.9 is adopted. Under this criterion, both residential types retain 11 indicators, and only the median resident consumption is removed because of its very high correlation with the income-related variable.

3.2 PCA-guided representation learning

Residential land efficiency is jointly shaped by multiple interrelated variables, and their relationships may be nonlinear. To balance interpretability and representational power, this study adopts a PCA-guided representation learning strategy.

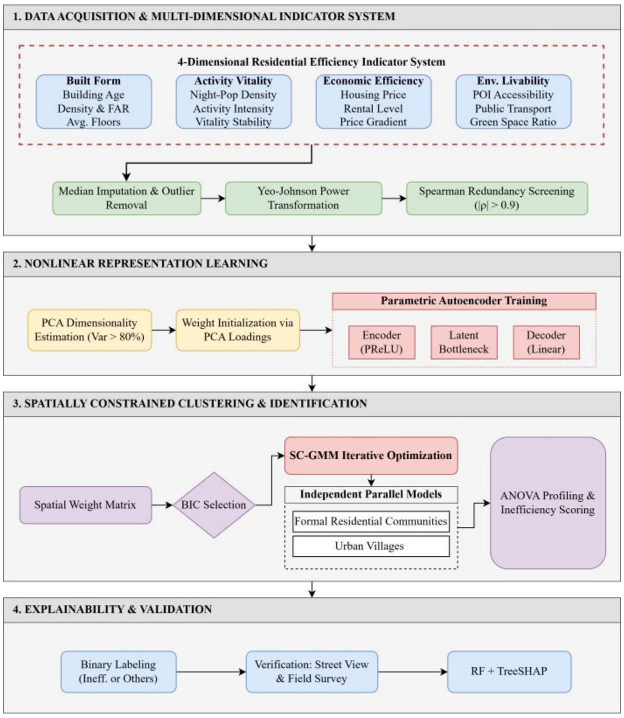

Figure 1. Overall methodological framework.

First, principal component analysis (PCA) is applied to the standardized indicator matrix for each residential type. The number of principal components explaining at least 80% of the cumulative variance is used to estimate the intrinsic dimensionality of the data. These retained components are not treated as the final analytical features; instead, they are used to determine the bottleneck size of a parametric autoencoder and to initialize the encoder weights.

Next, a parametric autoencoder is trained to learn a nonlinear latent representation of the original indicators. Compared with conventional PCA, the autoencoder can capture nonlinear interactions among variables. Compared with an unconstrained autoencoder, the PCA-guided initialization makes the learned latent space more stable and interpretable by anchoring it to the empirical linear structure of the data.

In the final implementation, the autoencoder adopts a single-hidden-layer encoder–decoder structure. The encoder maps the standardized input indicators to the latent space through a fully connected linear layer, followed by a learnable PReLU activation function, and the decoder reconstructs the input through a symmetric linear layer. The encoder and decoder weights are initialized using the retained PCA loading vectors, while all biases are initialized to zero. The network is trained with the Adam optimizer using a learning rate of 0.001 and a weight decay of $1 * 10^{-5}$. Mean squared error is used as the reconstruction loss. The model is trained for 300 epochs with a batch size of 256.

The purpose of this design is to preserve the empirical linear structure revealed by PCA while allowing the latent representation to adapt to nonlinear relationships through trainable activation and reconstruction learning.

3.3 Spatially constrained Gaussian mixture model

After representation learning, the low-dimensional latent features are used as inputs to a spatially constrained Gaussian mixture model (SC-GMM). GMM is a soft-clustering method that assumes the data are generated from a mixture of multiple Gaussian distributions. Unlike hard-clustering methods, it assigns each parcel a probability of membership across clusters, which makes it suitable for identifying transitional or mixed-efficiency residential areas.

To improve spatial coherence, spatial constraints are incorporated into the clustering process. These constraints encourage neighboring parcels with similar feature profiles to have more consistent cluster membership, thereby reducing unrealistic fragmentation in the clustering output. This is especially important in residential space, where adjacent parcels often share development history, service conditions, and neighborhood environments.

Model complexity is selected using the Bayesian Information Criterion (BIC), which balances model fit and parsimony. Formal residential communities and urban villages are modeled separately to account for their structural heterogeneity.

For the spatial constraint, a row-normalized sparse spatial-weight matrix is constructed for each residential type. Parcel adjacency is identified by polygon touching based on the parcel boundaries and a spatial index. In the SC-GMM, the spatial term is introduced through the weighted neighborhood membership matrix $W * R$, where W denotes the row-normalized spatial-weight matrix and R denotes the current soft-membership matrix. The final log-responsibility therefore combines the Gaussian likelihood term, the mixture-weight term, and a spatial regularization term controlled by β= 0.9.

In the citywide experiment, the number of clusters is first selected by BIC from a candidate range of 2~6 using an ordinary diagonal-covariance GMM, after which the SC-GMM is run with the selected cluster number for each residential type.

3.4 Identification of inefficient clusters

The SC-GMM produces multiple latent residential clusters for each residential type. However, the statistical output of clustering does not directly indicate which cluster represents inefficient land. Therefore, an additional interpretation step is required.

In this study, one-way analysis of variance (ANOVA) is used to compare the cluster-level means of the retained indicators. Clusters that show systematically poorer performance across multiple dimensions, such as lower vitality, weaker economic performance, poorer service accessibility, or unfavorable built-form characteristics, are interpreted as inefficient residential land. Rather than relying on a single indicator, the judgment is based on the combined profile of each cluster.

To distinguish inefficient clusters from merely distinct clusters, this study evaluates the standardized cluster means under a unified inefficiency-direction rule. Indicators whose larger values imply worse efficiency are treated as positive contributors to inefficiency, while indicators whose larger values imply better efficiency are directionally reversed. A cluster-level inefficiency score is then calculated by averaging these direction-adjusted values, together with dimension-specific scores for built form, activity vitality, economic efficiency, and environmental livability. The cluster with the highest overall inefficiency score is labeled as the inefficient type, and the ANOVA results are used as supporting evidence that the identified cluster is statistically distinct from the others.

3.5 Interpretation and validation

To enhance interpretability, the identified clustering outcomes are transformed into binary labels for post hoc analysis, distinguishing inefficient from non-inefficient parcels. A random forest classifier is then trained using the retained indicators as predictors. TreeSHAP is further employed to quantify the relative contribution of each variable and to reveal the factors most strongly associated with inefficient-cluster membership (Lundberg and Lee, 2017).

It should be noted that this step is used for post hoc interpretation of the clustering results rather than for causal inference. The identified variables therefore reflect associated factors rather than definitive determinants.

Validation is conducted at two complementary levels. First, an attribute back-check is used as an internal consistency test to examine whether the parcels identified as inefficient also show systematically weaker multidimensional profiles in the original attribute space. This step is not used as external ground truth, but as supporting evidence for the internal coherence of the clustering results.

Second, an external validation is conducted using street-view image interpretation and limited field observations. To control the validation workload while preserving representativeness, approximately 500 parcels are randomly sampled across the two residential types at the citywide scale. The sample is allocated proportionally by residential type, resulting in about 160 urban-village parcels and about 340 residential-community parcels. Within each residential type, stratified random sampling is further applied according to the proportions of parcels predicted as inefficient and non-inefficient, so that the sampled validation set remains close to the model-predicted class distribution of the full dataset.

For each sampled parcel, street-view imagery and, where available, field observations are used to determine whether it shows visible signs of inefficient residential conditions, such as aging building fabric, poor maintenance, weak public-space quality, inadequate service support, or poor accessibility. Based on these external labels, confusion matrices are constructed and overall accuracy and Cohen’s Kappa are calculated. To improve statistical transparency, 95% confidence intervals for the validation metrics can be estimated by bootstrap resampling.

4. Results

4.1 Indicator screening and dimensionality reduction

The initial candidate indicator system contains 12 indicators covering built form, activity vitality, economic efficiency, and environmental livability. After Spearman correlation screening, the indicator system remains largely stable for both residential types. Using a threshold of $|rho| > 0.9$, 11 indicators are retained for urban villages and formal residential communities. In both cases, only median resident consumption (MRC) is excluded because of its strong redundancy with median resident income (MRI).

PCA results show that a limited number of components can explain most of the variance in the original indicator system. Specifically, five principal components are retained for urban villages and six for formal residential communities, both exceeding the 80% cumulative variance threshold. These component numbers are then used to define the bottleneck dimensions of the PCA-guided autoencoder. The trained autoencoder further compresses the indicator system into a nonlinear latent space while preserving essential structural information for subsequent clustering.

4.2 Clustering and identification of inefficient residential land

The SC-GMM is applied separately to urban villages and formal residential communities using the learned latent representations. According to the BIC, the optimal number of clusters is six for urban villages and five for formal residential communities. These results indicate that the efficiency structure of the two residential types is differentiated and cannot be adequately represented by a single unified partition.

Cluster-profile analysis shows that one cluster in each residential type exhibits systematically weaker multidimensional performance and can therefore be interpreted as the inefficient type. For urban villages, cluster 3 is identified as the inefficient cluster, containing 388 parcels, which accounts for approximately 19.1% of all urban-village parcels in the citywide dataset. This type is characterized by older building stock, lower NPD and weaker relative neighborhood vitality (ECI_NPD), low rental and income performance, and substantially poorer accessibility to medical, educational, leisure, and transport services. Morphologically, it also shows lower development intensity than the more efficient village clusters. For formal residential communities, cluster 2 is identified as the inefficient cluster, containing 706 parcels, accounting for approximately 16.4% of all residential-community parcels.. This type is characterized by relatively high FAR and BD, lower NPD and lower ECI_NPD, and slightly weaker economic performance, indicating a combination of relatively intensive built form and comparatively weak actual use vitality.

The spatial distributions in Figure 2 show clear differences between the two residential types. Inefficient urban villages exhibit a more continuous and clustered pattern across the city, and many are distributed along the outer margins of larger village agglomerations. In contrast, inefficient formal residential communities are more scattered and fragmented, forming an island-like pattern within the broader urban fabric. These contrasting spatial signatures support the need for differentiated regeneration strategies.

4.3 Validation results

External validation based on street-view interpretation and limited field observation indicates that the proposed framework achieves acceptable identification performance for both residential types. As reported in Table 3, the overall accuracy reaches 0.850 for formal residential communities and 0.881 for urban villages, while the corresponding Kappa coefficients are 0.641 and 0.732, respectively.

These results suggest that the identification outcome is broadly consistent with observed residential conditions. In particular, the higher Kappa value for urban villages indicates stronger agreement beyond chance, implying that inefficient urban villages may exhibit more visually recognizable spatial features than inefficient formal residential communities.

https://doi.org/10.5194/isprs-archives-XLIX-B4-2026-167-2026 | © Author(s) 2026. CC BY 4.0 License.

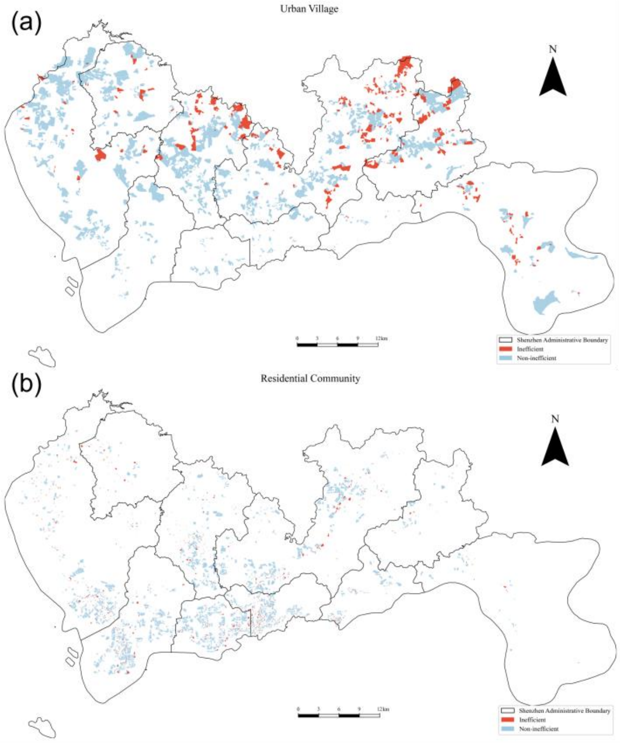

Figure 2. Spatial distribution of (a) inefficient urban villages and (b) inefficient formal residential communities in Shenzhen.

Actual \ Predicted	inefficient formal residential communities		inefficient urban villages	
	Positive	Negative	Positive	Negative
Positive	75	36	43	14
Negative	15	214	5	98
Accuracy	0.850		0.881	
Kappa	0.641		0.732	

Table 3. Confusion matrix, accuracy, and Kappa.

4.4 Associated factors of inefficient residential land

The random forest and TreeSHAP results in Figure 3 reveal that the factors associated with inefficient-cluster membership differ substantially between urban villages and formal residential communities. For urban villages, the most influential variables are primarily related to environmental livability and local socioeconomic conditions. Medical service accessibility, leisure and entertainment accessibility, public transport accessibility, median resident income, and educational service accessibility all rank among the strongest associated factors. The SHAP patterns further indicate that lower service accessibility and weaker socioeconomic support are closely associated with the probability of belonging to the inefficient urban-village cluster, while building age also contributes to inefficiency.

For formal residential communities, the dominant associated factors are activity vitality and development intensity. NPD and the ECI_NPD are the most important variables, followed by building density and floor area ratio. The SHAP distributions indicate that lower NPD, lower ECI_NPD, and higher development intensity are associated with inefficient-cluster membership, while building age and weaker market performance also play supplementary roles.

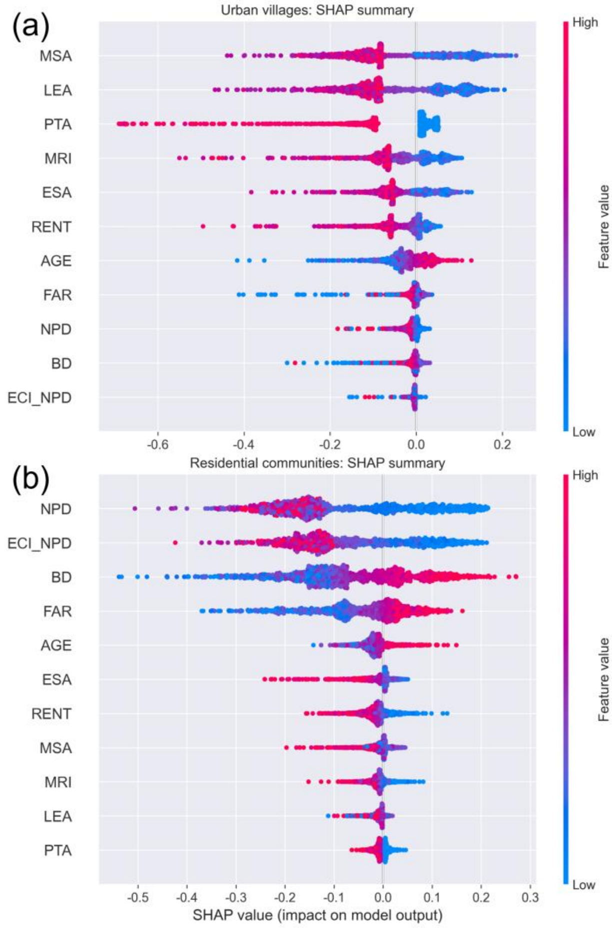

Figure 3. TreeSHAP-based feature importance and summary plots for (a) urban villages and (b) formal residential communities.

Taken together, these findings suggest that inefficient urban villages are more strongly shaped by deficiencies in surrounding service support and local socioeconomic conditions, whereas inefficient formal residential communities are more closely associated with a mismatch between built intensity and actual residential vitality. This contrast further supports the view that inefficient residential land should not be treated as a homogeneous planning target.

5. Conclusions

5.1 Data and methodological implications

This study contributes to the identification of inefficient residential land in three ways.

First, the framework incorporates mobile-phone signaling data as a dynamic source of human-activity information. Compared with conventional static demographic or land-use statistics, mobile-phone signaling data provide a finer-grained spatiotemporal representation of population distribution and locational behavior, making it possible to capture actual residential use intensity and neighborhood vitality more directly. When combined with parcel-based built-form indicators, rental information, and POI-derived service accessibility, such data help reveal the mismatch between physical development and real human activity patterns, thereby improving the sensitivity of inefficient residential land identification.

Second, it treats formal residential communities and urban villages as distinct analytical objects, thereby explicitly incorporating residential heterogeneity into the evaluation process. This is important because the two residential forms differ not only in physical morphology but also in occupancy patterns, market logic, and service dependency.

Third, the proposed framework combines representation learning with probabilistic spatial clustering. The PCA-guided autoencoder helps retain essential structure while allowing nonlinear relationships to be captured, and the SC-GMM provides a more flexible clustering strategy than hard partitioning. This is particularly useful for residential land, where transitional parcels and mixed-efficiency areas are common. In addition, the interpretation of Gaussian-mixture-based partitions requires care because statistical mixture components do not automatically correspond to substantively meaningful clusters, which further motivates the cluster-profile interpretation adopted in this study (Hennig, 2010).

5.2 Planning implications

The empirical findings suggest that inefficient formal residential communities and inefficient urban villages exhibit different spatial structures and likely require different regeneration approaches. The scattered and island-like distribution of inefficient formal residential communities indicates that they may be better addressed through selective renewal, micro-regeneration, and facility upgrading at the neighborhood scale. In contrast, the more contiguous pattern of inefficient urban villages implies stronger spatial dependence and may justify more coordinated regeneration in clustered areas.

In addition, the associated-factor analysis suggests that regeneration should move beyond a purely morphological understanding of inefficiency. For formal residential communities, lower average nighttime population density and weaker relative neighborhood vitality, combined with relatively high development intensity appears to be a key signal of inefficient use. For urban villages, deficiencies in medical, educational, transport, and leisure accessibility, together with weaker local socioeconomic conditions, are more strongly associated with inefficiency. A differentiated strategy is therefore necessary.

5.3 Limitations and future work

This study still has several limitations. Although multiple datasets are integrated, the current framework remains constrained by both data timeliness and representational breadth. Future research could incorporate more up-to-date and diversified data sources, such as continuous population-activity records, street-view imagery, dynamic housing-listing information, and urban-renewal project data, in order to enhance the fine-grained characterization of inefficient residential space. Such extensions would help improve the joint representation of residential environmental quality, spatial vitality, and renewal status, especially in terms of temporal consistency across multiple dimensions.

In addition, the present study mainly focuses on the identification of current inefficient residential land patterns. Future work could further integrate renewal-process information and policy implementation contexts to identify the evolutionary trajectories, stage characteristics, and dynamic response processes of inefficient residential space. This would help move the research from static identification toward stronger decision support, and strengthen its links with practical scenarios such as urban regeneration, housing provision, public-service allocation, and area-based governance. It would also support the exploration of differentiated governance pathways and optimization strategies for different types of inefficient residential land.

Acknowledgements

This research was funded by National Key Rsearch and Development Project of China (Grant No. 2023YFC3804805) , the Open Fund of Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources (Grant No. KF-2022-07-008), Tencent “Rhinoceros Birds” - Scientific Research Foundation for Young Teachers of Shenzhen University and Shenzhen Science and Technology Program (No. 20231121100546001).

References

Chen, K., Long, H., Liao, L., Tu, S., Li, T., 2020: Land use transitions and urban-rural integrated development: Theoretical framework and China’s evidence. Land Use Policy, 92, 104465.

Dong, X., Zhang, X., Zhou, Q., 2023: Underload and overload communities: Revealing the conflicts between population distribution and carrying capacity at an inner-city community scale. Sustainable Cities and Society, 98, 104793.

Du, S., Zhang, X., Lei, Y., Huang, X., Tu, W., Liu, B., Du, S., 2024: Mapping urban functional zones with remote sensing and geospatial big data: A systematic review. GIScience & Remote Sensing, 61(1), 2404900.

Estoque, R.C., Ooba, M., et al., 2021: Monitoring global land-use efficiency in the context of the UN 2030 Agenda for Sustainable Development. Habitat International, 115, 102403.

Hennig, C., 2010: Methods for merging Gaussian mixture components. Advances in Data Analysis and Classification, 4, 3–34.

Herzig, A., Nguyen, T.H., et al., 2018: Assessing resource-use efficiency of land use. Environmental Modelling & Software, 107, 34–49.

Jin, R., Huang, C., Wang, P., Ma, J., Wan, Y., 2023: Identification of inefficient urban land for urban regeneration considering land use differentiation. Land, 12(10), 1957.

Liu, S., Xiao, W., Li, L., Ye, Y., Song, X., 2020: Urban land use efficiency and improvement potential in China: A stochastic frontier analysis. Land Use Policy, 99, 105046.

Long, H., 2014: Land use policy in China: Introduction. Land Use Policy, 40, 1–5.

Lundberg, S.M., Lee, S.-I., 2017: A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

Qi, Y., Zhao, Y., Guo, J., Wang, Y., 2025: Identification and redevelopment of inefficient industrial land in resource-exhausted cities: A case study of Hegang, China. Land, 14(6), 1292.

Qiao, H., Huang, J., 2023: Study on inefficient land use determination method for cities and towns from a city examination perspective. Applied Mathematics and Nonlinear Sciences, 8(1), 2425–2438.

Schiavina, M., Melchiorri, M., et al., 2022: Land use efficiency of functional urban areas: Global pattern and evolution of development trajectories. Habitat International, 123, 102543.

Sun, Y., Hu, H., Han, Y., Wang, Z., Zheng, X., 2023: Large-scale automatic identification of industrial vacant land. ISPRS International Journal of Geo-Information, 12(10), 409.

Wang, H., Liu, Y., Sun, L., Ning, X., Li, G., 2025: Assessment of Chinese urban land-use efficiency (SDG 11.3.1) utilizing high-precision urban built-up area data. Geography and Sustainability, 6(1), 100210.

Wang, X., Bao, X., Ge, Z., Xi, J., Zhao, Y., 2024: Identification and redevelopment of inefficient residential landuse in urban areas: A case study of Ring Expressway Area in Harbin City of China. Land, 13(8), 1238.

https://doi.org/10.5194/isprs-archives-XLIX-B4-2026-167-2026