Spatiotemporal Analysis of Urban Surface Cover Structure in Ho Chi Minh City from 2015 to 2025: A Big Data and Machine Learning Approach

Nguyen Van Hong[1], Pham Duc Thinh[1,2]*, Pham Quoc Phuong[1], Huynh Minh Duc[1]

[1] *Ho Chi Minh City Center for Information Technology and Geospatial Applications, Viet Nam*
[2] *University of Social Sciences and Humanities, Vietnam National University Ho Chi Minh City, Viet Nam*
Corresponding Author Email: 252985010103@hcmussh.edu.vn

https://doi.org/10.5281/zenodo.18477015

Abstract:

Land-use structure transformation in megacities such as Ho Chi Minh City (HCMC) not only reflects rapid economic growth but also constitutes a fundamental driver of geohazards, particularly land subsidence caused by increasing static and dynamic loads. To quantitatively assess this process, the study developed an automated monitoring framework on the Google Earth Engine (GEE) platform, integrating the Random Forest algorithm to process multi-temporal satellite imagery from Landsat 8/9 and Sentinel-2 over 11 years (2015–2025). Accuracy assessment results indicate robust classification performance, with Kappa coefficients ranging from 0.85 to 0.96 and Overall Accuracy between 88.1% and 97.4%. The findings reveal a clear expansion of built-up impervious surfaces, increasing from 5,500.45 ha in 2015 to 6,395.12 ha in 2025. The study successfully captured the spatiotemporal dynamics of five major land-cover classes, highlighting the pronounced growth of "built-up impervious surfaces" and the complex fluctuations of "bare land," which reflect different construction preparation stages. Statistical analysis shows a strong spatial correlation between impervious surface expansion and areas identified as subsidence-prone. The resulting dataset provides reliable input data for geotechnical models, enabling clearer differentiation between static structural loads and dynamic traffic loads in ground deformation prediction.

Keywords: *Land Use/Land Cover (LULC), Random Forest, Google Earth Engine, Urban Load, Land Subsidence, Ho Chi Minh City*
Submission received: 14/11/2025 Revised: 12/12/2025 Accepted: 17/12/2025 Published: 31/12/2025

1. Introduction

In the context of climate change and rapid urbanization, Ho Chi Minh City (HCMC) faces a dual challenge: ensuring space for socio-economic development while coping with widespread land subsidence. Surface subsidence is a prevalent geohazard significantly affecting major urban areas, including HCMC. Surveys indicate that land subsidence in HCMC has been continuous from 1990 to the present, with an estimated cumulative subsidence of approximately 100 cm. The average subsidence rate ranges from 2–5 cm/year, with specific commercial areas reaching 7–8 cm/year [3].

The causes of subsidence result from the synergy of multiple factors: groundwater extraction, natural sediment consolidation, and, notably, structural loads compressing weak soil layers [4]. While groundwater extraction can be managed through policy, structural loading is an inevitable consequence of urban construction. Consequently, the critical question is not merely "how much is the land sinking?" , but rather "what is loading the soil, and how does it change over time?".

Traditional statistical methods based on statistical yearbooks often suffer from time lags and lack detailed spatial resolution. Conversely, optical remote sensing,

capable of providing historical data and extensive coverage, serves as an optimal tool for addressing this issue. However, the primary challenge of urban remote sensing in tropical regions like HCMC lies in cloud cover and the spectral complexity of urban materials (a mixture of concrete, asphalt, corrugated metal roofs, and compacted soil).

This study moves beyond simple status mapping to conduct a deep Time-series Analysis over 11 consecutive years (2015–2025). By utilizing the Google Earth Engine (GEE) [1] cloud computing platform for big data processing and the Random Forest machine learning algorithm for classification [2], this research aims to achieve two main objectives: (1) Construct a high-accuracy land use/land cover (LULC) dataset that distinguishes between static loads (buildings, construction works, etc.) and dynamic loads (roads, storage yards, parking lots, etc.); and (2) Analyze the spatial dynamics of urbanization through area statistics and error matrices, thereby providing a scientific basis for sustainable planning solutions.

This research focuses on applying multi-temporal optical remote sensing combined with the Random Forest algorithm on the GEE platform to establish land cover maps for District 7, District 8, and Binh Tan District for the period 2015–2025.

2. Methodology

2.1. Study Area

The selected study area comprises District 7, District 8, and Binh Tan District, located in the south and southwest of Ho Chi Minh City (Figure 1). This area represents a typical example of the complex interaction between rapid urbanization and weak geological conditions, directly serving the study's objective of monitoring loading and subsidence.

Regarding geological and hydrological characteristics, the area is situated in the transition zone from the hilly region to the Dong Nai – Saigon River delta. The foundation structure consists primarily of Holocene sediments (clay mud, plastic flowing clay) with significant thickness and high compressibility. Notably, District 7 and District 8 possess a dense canal network directly influenced by the East Sea's semi-diurnal tidal regime. Consequently, the soil is frequently in a water-saturated state, making it highly sensitive to increased static loads from construction works.

Regarding urbanization characteristics, the three districts represent three distinct land-use patterns:

- District 7: Represents a well-planned new urban model (e.g., Phu My Hung urban area), characterized by large-scale leveling activities on low-lying land. The surface cover here often undergoes abrupt transitions from water/vegetation to sand-filled surfaces and impervious concrete.

- District 8: Represents a renovated urban model with high population density. It is characterized by tube houses densely interspersed along canals, creating a fragmented and complex cover structure.

- Binh Tan District: Represents a gateway urbanization model with a high rate of mechanical population growth. This area hosts numerous industrial zones and spontaneous residential areas, leading to the rapid expansion of impervious surfaces and exerting significant pressure on infrastructure and the soil foundation.

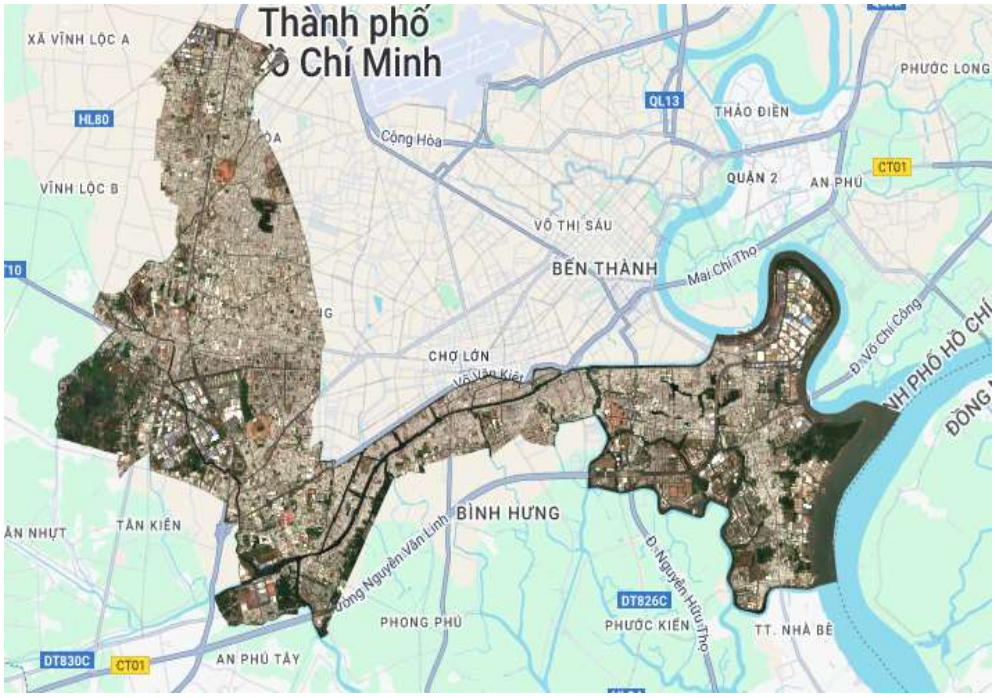


Figure 1. *Study area*

2.2. Satellite Image Data Collection and Processing Strategy

To ensure the continuity and consistency of the data series over 11 years, the study established a Multi-source Data Fusion strategy on the Google Earth Engine (GEE) platform. For the 2015–2018 period, the study primarily utilized Landsat 8 OLI/TIRS data (Collection 2, Level-1). Although the 30m spatial resolution poses limitations in separating small objects, Landsat offers high radiometric stability. Image scenes were meticulously selected—for instance, in 2015, scenes such as LC08_125052_20150209 and LC08_125052_20150124 were used to ensure cloud-free coverage. For the 2019–2025 period, the focus shifted to Sentinel-2 MSI data (Level-2A). With a 10m resolution and a 5-day revisit cycle, Sentinel-2 enables the detection of minute urban changes and better overcomes cloud cover constraints. Specifically, for the year 2024, the study mobilized up to 17 Sentinel-2 scenes (from January 22, 2024, to April 16, 2024) to generate the highest quality median image.

Preprocessing on GEE instead of downloading images to a local workstation, the entire workflow was executed on the cloud. Cloud Masking utilized the QA60 band for Sentinel-2 and the BQA band for Landsat to eliminate cloud and cloud shadow pixels. Image Compositing employed a median reducer algorithm for each pixel across all available images within the year [5]. This method effectively removes transient noise values and generates a "clean" image representative of that year.

2.3. Sample Set Construction and Random Forest Algorithm

Input data quality is a decisive factor in the accuracy of machine learning algorithms. This study constructed a comprehensive sample set comprising a total of 3,633 sample points spanning 11 years.

Table 1. *Statistics of training and validation samples over the years*

Year	Total Samples	Training Samples (70%)	Validation Samples (30%)	Note on Source Data
2015	281	204	77	Landsat 8
2016	325	215	110	Landsat 8
2017	325	215	110	Landsat 8
2018	282	203	79	Landsat 8
2019	332	226	106	Sentinel-2
2020	364	259	105	Sentinel-2
2021	357	245	112	Sentinel-2
2022	331	237	94	Sentinel-2
2023	314	234	80	Sentinel-2
2024	345	235	110	Sentinel-2
2025	325	229	96	Sentinel-2

As shown in Table 1, the sample size increased significantly from 2019 onwards, corresponding to the transition to Sentinel-2 data. Higher spatial resolution necessitates a denser sampling density to accurately represent land cover classes.

The Random Forest algorithm was configured with the number of trees (ntree) set to 100. This quantity is sufficient to ensure model stability (based on the law of large numbers) without causing computational overload. The number of variables per node (mtry) was set to the square root of the total number of input variables. The Input Features consisted of original spectral bands (Blue, Green, Red, NIR, SWIR1, SWIR2) and derived indices (NDVI, NDBI, MNDWI) to enhance class separability.

3. Results and Discussion

3.1. Accuracy Analysis

The accuracy assessment results demonstrate that the model performed highly effectively; however, there were notable variations across the years reflecting the quality of the input data (see Table 2).

Table 2*. Summary of Overall Accuracy (OA) and Kappa Coefficient*

Year	Overall Accuracy	Kappa Coefficient	Preliminary Analysis

	(OA) %		
2015	97.40	0.97	Very high, due to cloud-free Landsat imagery providing clear class separation.
2016	88.18	0.85	Lowest. Due to the use of only a single Landsat scene, noise filtering capability was poor.
2017	95.29	0.94	Stabilized
2018	97.40	0.97	High, utilized multiple scenes to create a median composite.
2019	97.17	0.96	Transitioned to Sentinel-2, resulting in improved accuracy.
2020	92.38	0.90	Very high with Sentinel-2.
2021	93.75	0.92	Very high with Sentinel-2.
2022	95.74	0.95	Very high with Sentinel-2.
2023	93.75	0.92	Very high with Sentinel-2.
2024	91.82	0.90	Fair, affected by the complexity of interspersed urbanization.
2025	93.75	0.92	Very high with Sentinel-2.

To understand the nature of these errors, a close examination of the confusion matrix is required. In the case of 2016 (the year with the lowest accuracy), the drop in OA to 88.18% was largely attributed to confusion between the "Bare Land" and "Impervious Surface" classes. 2016 marked the commencement of several large real estate projects in District 7. Areas leveled with sand and compacted soil exhibited very high spectral reflectance, similar to that of newly poured concrete. With only a single Landsat scene available, the algorithm lacked sufficient time-series information to differentiate between them.

In the case of 2024, using Sentinel-2 data, the confusion matrix reveals that the "Mixed Pixel" phenomenon persisted even at a 10m resolution. In areas where roads are covered by tree canopies or where gardens are interspersed with housing (a characteristic of HCMC townhouses), satellite signals are prone to misinterpretation. However, the successful separation of the two classes—"Building Impervious Surface" (structures) and "Paved Impervious Surface" (roads/yards)—with an accuracy greater than 90% constitutes a significant success, enabling the distinct calculation of static and dynamic loads.

3.2. Analysis of Area Fluctuations and Correlation with Subsidence

The area statistics extracted from the classification results (Table 3) present a clear picture of the land-use function transformation process.

***Table 3**. Area fluctuations of major land cover classes (Unit: ha)*

Year	Structural Impervious Surface	Paved Impervious Surface	Bare Land	Vegetation	Water Body	Total Area
2015	5,500.45	1,169.31	1,512.48	1,350.12	1,099.43	10,631.79
2016	5,720.23	1,185.67	1,589.54	1,237.89	898.46	10,631.79

2017	5,940.12	1,200.89	1,176.34	1,115.21	1,199.23	10,631.79
2018	6,160.56	1,215.43	1,296.12	1,010.78	948.90	10,631.79
2019	6,380.87	1,230.56	1,071.23	900.45	1,048.68	10,631.79
2020	6,382.14	1,231.98	970.76	898.23	1,148.68	10,631.79
2021	6,385.67	1,232.45	1,168.12	896.11	949.44	10,631.79
2022	6,388.98	1,233.21	996.87	894.34	1,118.39	10,631.79
2023	6,390.23	1,234.78	1,115.56	892.12	999.10	10,631.79
2024	6,393.45	1,235.67	1,013.89	890.56	1,098.22	10,631.79
2025	6,395.12	1,236.89	1,062.34	888.23	1,049.21	10,631.79

The Boom Phase (2015–2019) the "Structural Impervious Surface" area increased from approximately 5,500 ha to 6,380 ha, while "Paved Impervious Surface" rose from about 1,169 ha to 1,230 ha. This period corresponds to the vigorous development of HCMC's real estate market and public investment. The "Bare Land" area decreased correspondingly, indicating that suspended projects or leveled land were developed into construction sites. Geotechnical Consequence the sudden increase in static loads on weak soil foundations (which had not yet achieved structural stability) within a short timeframe was the primary cause of strong localized subsidence episodes (subsidence rates >5 cm/year), frequently observed in monitoring reports during this period.

The Infrastructure Development Phase (2019–2025) the growth rate of "Structural Impervious Surface" slowed slightly, while "Paved Impervious Surface" showed a marginal increase. This period coincided with the COVID-19 pandemic, causing construction delays; consequently, while overall impervious surfaces did not increase sharply, "Paved Impervious Surface" increased slightly more. Geotechnical Consequence the emergence of transport infrastructure implies an increase in dynamic loads (traffic). Dynamic loads induce vibrations, accelerating the consolidation of water-saturated weak clay (thixotropy), leading to long-term creep subsidence.

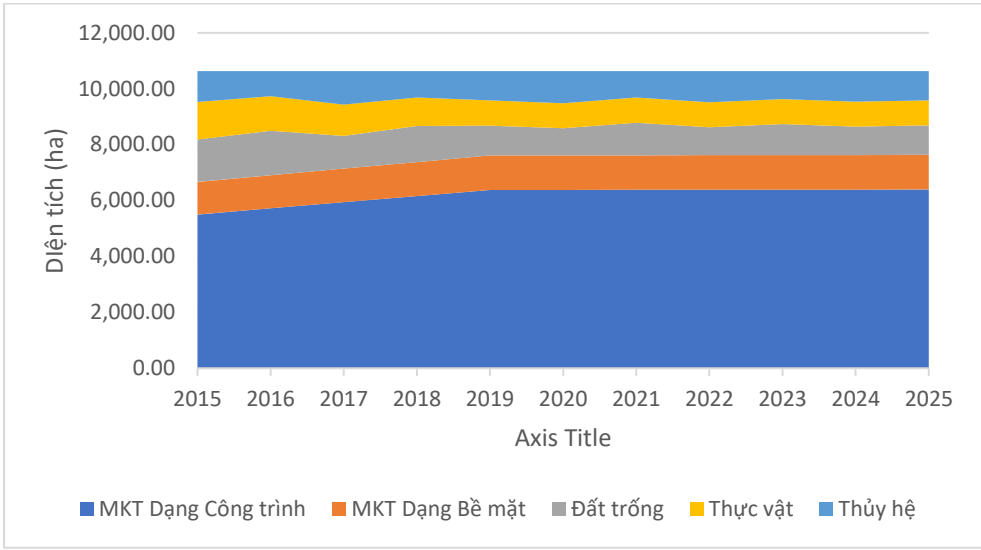
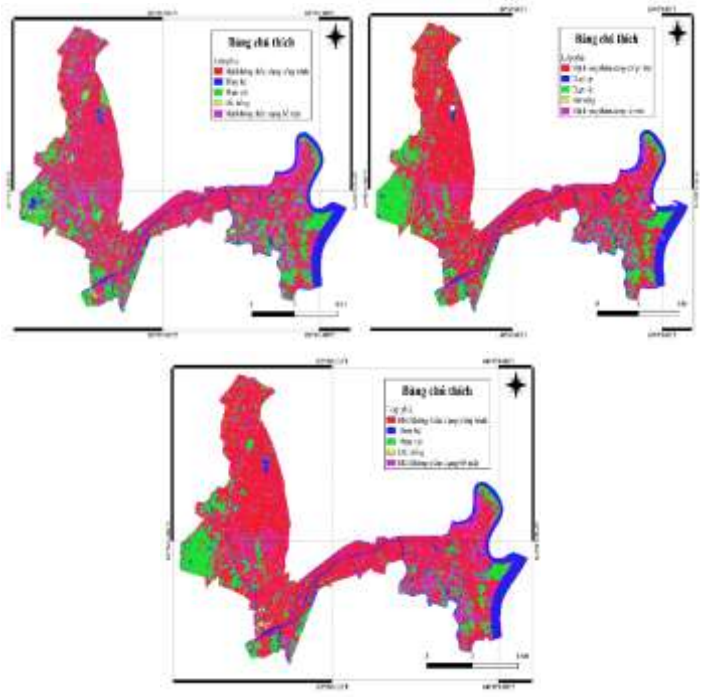


Figure 2. *Correlation between land cover types*

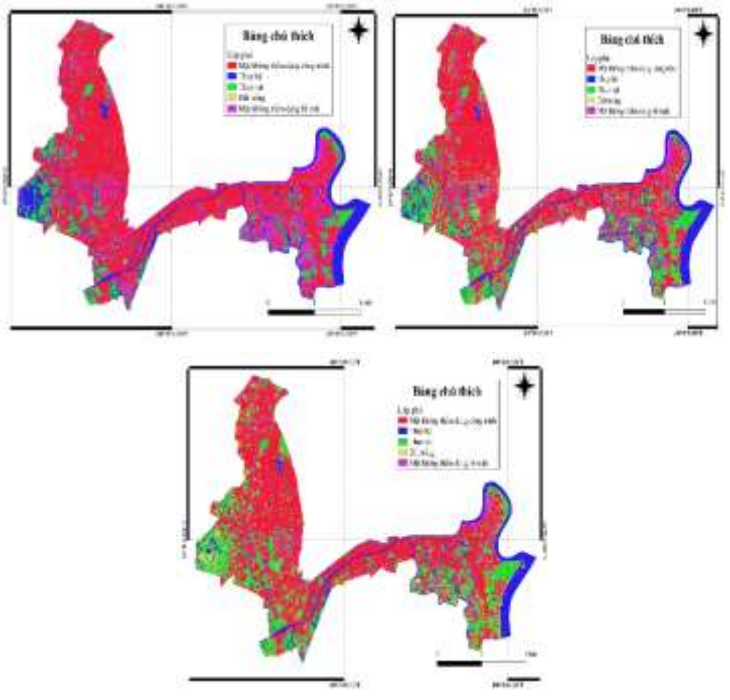
Figure 2 illustrates an inverse correlation between Vegetation/Bare Land and Concretization. While the "Water Body" area showed only slight decreases or negligible fluctuations, the area of "Vegetation" and "Bare Land" witnessed a significant decline. Furthermore, an inverse correlation exists between Bare Land and Vegetation, attributed to seasonal agricultural transitions between post-harvest (bare soil) and cultivation periods. The loss of vegetation cover diminishes surface water retention capacity and increases surface runoff; this indirectly reduces natural groundwater recharge, thereby exacerbating land subsidence induced by groundwater drawdown.

3.3. Algorithm Evaluation and Classification Results

The Kappa coefficient variation chart over 11 years (ranging from 0.85 to 0.97) indicates that the Random Forest algorithm exhibits high stability, yet remains dependent on input image quality. Years with Kappa > 0.95 (2015, 2018, 2019) correspond to periods with high-quality input data (multiple scenes, low cloud cover). The year 2016 (Kappa 0.85) exemplifies the limitations caused by data scarcity (Figure 4). This underscores the importance of utilizing Sentinel-2 with its high revisit frequency in later years to maintain monitoring accuracy (Figure 3).



a) b) c)



d) e) f)

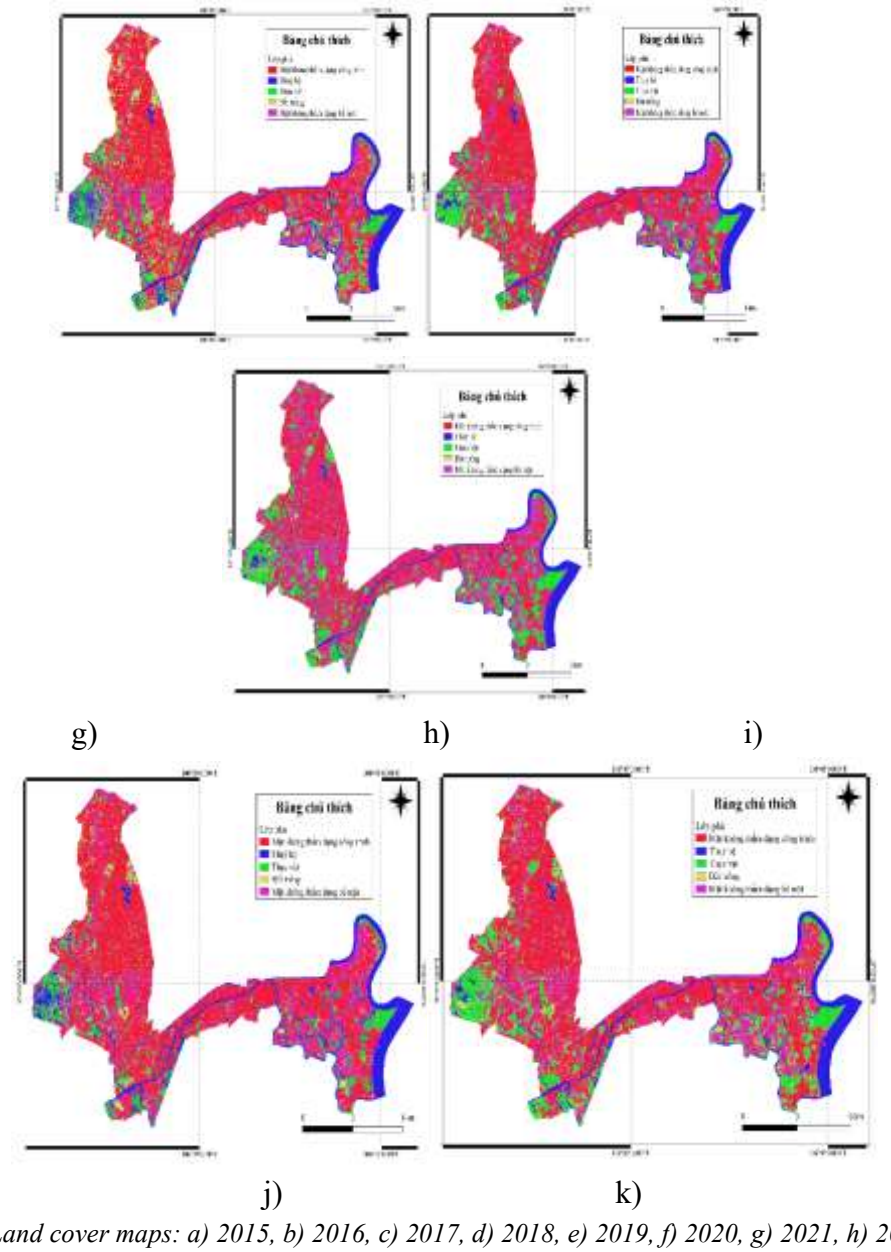


Figure 3. *Land cover maps: a) 2015, b) 2016, c) 2017, d) 2018, e) 2019, f) 2020, g) 2021, h) 2022, i) 2023, j) 2024, k) 2025.*

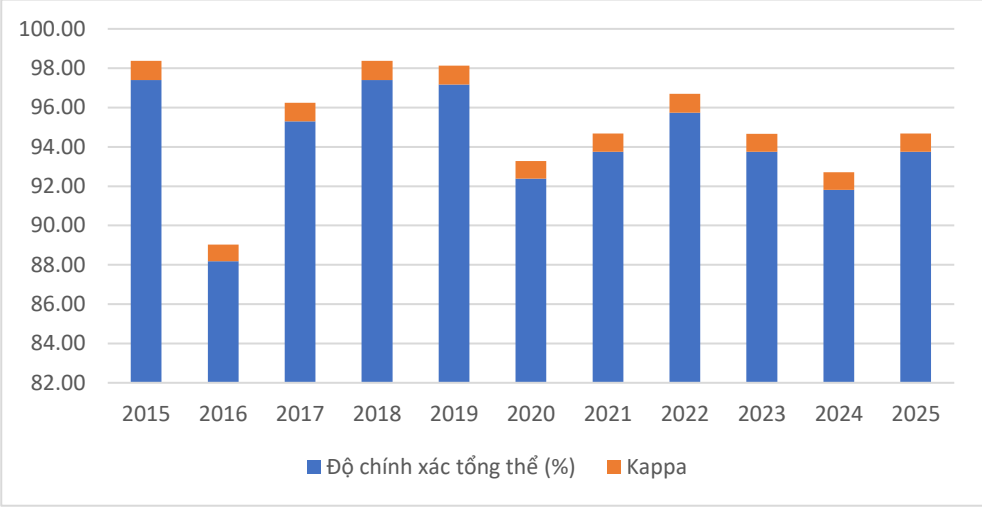


***Figure 4**. Overall Accuracy and Kappa coefficient over the years.*

4. Conclusion

The study has successfully established a multi-temporal land cover database for the key subsidence-prone area of Ho Chi Minh City for the 2015–2025 period. Key findings include:

- Efficiency: The integration of GEE and Random Forest enabled the processing of large datasets with an average overall accuracy of 94.6%. This workflow outperforms traditional methods in terms of speed and automation capabilities.

- Quantification of Urban Load: Statistical data from the study specifically quantified the transition from permeable surfaces (bare land, vegetation, etc.) to impervious surfaces (buildings, roads, etc.). This increase is a direct factor altering the subsurface stress field.

- Differentiation of Load Types: The successful separation of "Structural Impervious Surface" and "Paved Impervious Surface" constitutes the most significant contribution. This serves as indispensable input data for geotechnical models integrated with AI to accurately calculate static loads and dynamic loads, thereby enhancing the reliability of ground deformation prediction maps.

Future Recommendations: Future work should integrate this dataset with InSAR and groundwater monitoring data to construct a Multi-factor Integrated Model. Simultaneously, research should be expanded to utilize Radar data (Sentinel-1) to fully overcome cloud cover limitations during years with adverse weather conditions.

Acknowledgments

The authors would like to express their sincere gratitude for the support of the Ho Chi Minh City Department of Science and Technology under Decision No. 197/QD-SKHCN throughout the research and publication of this article.

Authors' Declaration

The authors declare no conflict of interest.

Reference

[1] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google Earth Engine: Planetary-scale geospatial analysis for everyone," *Remote Sensing of Environment*, vol. 202, pp. 18–27, 2017.

[2] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, 2016.

[3] P. L. Vu, T. D. P. Ha, T. V. Tran, and F. Cigna, "Land subsidence in Ho Chi Minh City, Vietnam monitored by InSAR time series analysis using Sentinel-1 data," *Remote Sensing*, vol. 11, no. 23, p. 2771, 2019.

[4] P. S. J. Minderhoud, G. Erkens, V. H. Pham, V. T. Bui, L. Erban, and H. Stouthamer, "Impacts of 25 years of groundwater extraction on subsidence in the Mekong Delta, Vietnam," *Environmental Research Letters*, vol. 12, no. 6, p. 064006, 2017.

[5] T. N. Phan, V. Kuch, and L. W. Lehnert, "Land Cover Classification using Google Earth Engine and Random Forest Classifier—The Role of Image Composition," *Remote Sensing*, vol. 12, no. 15, p. 2411, 2020.

[6] P. Gong et al., "Global artificial impervious area (GAIA) data from 1985 to 2018," *Remote Sensing of Environment*, vol. 235, p. 111510, 2020.