https://doi.org/10.48047/AFJBS.6.13.2024. 264-275

African Journal of Biological Sciences

Application of machine learning model and cloud computing for urban heat island forecasting in Hanoi city

*Dong Phuong Nguyen**

Faculty of Environment, Hanoi University of Mining and Geology;

E-Mail: nguyenphuongdong@ humg.edu.vn

Mai Hoa Thi Phan

Faculty of Environment, Hanoi University of Mining and Geology;

E-Mail: phanthimaihoa@humg.edu.vn

***Corresponding Author:** *Dong Phuong Nguyen, Email: phuongdongmdc@gmail.com*

Article History
Volume 6, Issue 13, 2024
Received: 18June 2024
Accepted: 02July 2024
doi:*10.48047/AFJBS.6.13.2024.264-275*

Abstract

Urban heat islands (UHI) pose significant environmental and health challenges in rapidly urbanizing cities like Hanoi. This study presents an integrated approach utilizing machinelearning model and cloud computing to predict and delineate UHIs in Hanoi. Leveraging high-resolution satellite imagery, meteorological data, population dentistry data and land cover factors, we develop a robust predictive model that accurately identifies UHI- prone areas.

The data preprocessing involves normalizing diverse data sources, handling missing values, and ensuring spatial-temporal alignment. Feature selection is performed to identify the most influential factors contributing to UHI, including land surface temperature, vegetation index, population density, and urban morphology. Results indicate that the machine learning models, particularly the ensemble methods, exhibit high predictive accuracy, with the Random Forest model achieving an R^2 of 0,66. In conclusion, applying machine learning models and cloud computing presents a powerful framework for predicting and managing urban heat islands. The study's innovative approach enhances understanding of UHI phenomena in Hanoi and provides a scalable solution adaptable to other urban settings. Future research should focus on refining model accuracy by incorporating additional data sources and exploring the socio-economic impacts of UHI mitigation strategies.

Keywords: Urban heat island, machine learning, cloud computing

1. Introduction

Climate change is one of the threats increasingly affecting humanity around the world. Vietnam is also one of the countries hit hard by climate change. Over the past 50 years, the average temperature in the country has increased by about 0,62 ^oC , especially with the increase in hot days in summer in recent years. This risk affects economic development and sustainability goals, impacts human health, and increases economic and energy damage and losses. An urban heat island is a phenomenon typically observed in large cities and developed urban areas with high population and development densities, but due to the impact of climate change, the effects of this phenomenon are becoming more common and more severe.

Due to the peculiarities of climatic parameters, including the temperature regime in the territory, it is often constantly changing and is influenced by many external factors, so direct measurement methods face many difficulties due to the declaration of implementation costs, area, and network of monitoring points. Remote sensing imaging technology is one of the practical solutions that help continuously monitor and track changes in air and surface temperature and update changes quickly [5-10, 16]. Using remote sensing image data, we can calculate, estimate, and separate the magnitude of the urban heat island effect [4].

However, the results of the above studies show that interpreting spectral channel images based on the differences in reflectance characteristics and spectral bands in each channel allows users to extract the required layer of information, which serves for standard zoning, monitoring, and forecasting [3]. However, in these studies, remote sensing images are used only to calculate specific indicators and cannot show a clear correlation between the change in surface temperature and other factors such as rainfall and population, land cover, etc. Therefore, this study aims to present the interpretation method data from the images remote sensing dataset combined with the random forest classification (RF) method on the Google Earth Engine cloud computing platform to estimate and predict urban heat island zoning in Hanoi city based on land cover, rainfall, population, elevation, and land surface temperature.

2. Materials and research methods

Research area

The research area is the entire boundary of Hanoi city within the range from 20°34' to 21°18' North latitude and from 105°17' to 106°02' East longitude. The terrain of Hanoi city is characterized by a gradual trend from north to south and from west to east, with diverse terrain including high hills and mountains in the north and west, as well as plains accounting for three-quarters of the natural area. The city is densely populated and convenient for construction, economic trade, and industrial development.

Hanoi city was chosen for research because of its dense population, high level of resident activity, high construction density, and sparse, declining flora [14]. The city's surface is covered mainly with construction materials such as brick, steel, concrete, and asphalt, which have the characteristics of a tight, impermeable, and dark surface that increase the ability to absorb light energy and convert it into heat, thus further exacerbating the urban heat island phenomenon [15]. In recent years, climate change has affected Hanoi, so in the summer, there are intense heat waves for a week with temperatures up to 42,5 ^oC.

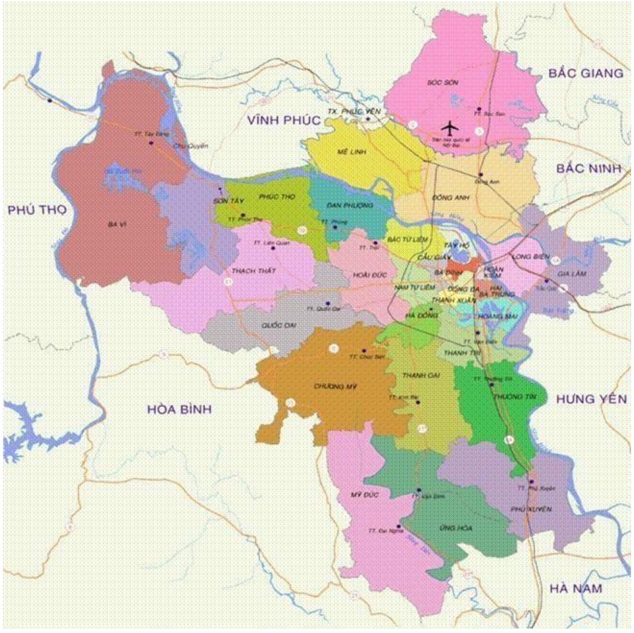

Figure 1. Boundaries and location of the study area

Data and Research methods

a. Database

Land surface temperature (LST) is an important parameter in urban heat island estimation and zoning studies. As mentioned above, assessing temperature changes in urban areas using traditional monitoring methods faces many limitations related to equipment, human resources, economics, and the requirements of long monitoring periods. With the outstanding advantages of diversity, continuous updating, long data series, and high accuracy, remote sensing technology has been widely applied and proven highly effective and reliable in surface temperature change estimation research. There have been many different studies that have used satellite image sources such as MODIS imagery, Landsat imagery, or Sentinel imagery to estimate temperature change in large urban areas [5-10] or the correlation between surface temperatures and differences in land cover layers [5, 7, 10]. This study uses primary sources, including MODIS image datasets, to calculate land surface temperature median values. The MODIS satellite image data provides monthly average land surface temperatures for 2000-2017.

Additionally, studies using datasets as selection parameters for machine learning include:

- Terrain data used according to the NASADEM dataset is cited in GEE under information ee.Image("NASA/NASADEM_HGT/001"). It is a broadband model with a global resolution of 30 m, providing one of the high-precision datasets suitable for scientific research in geography, hydrology, geology, and environmental management. [11]

- Population data in the study is used from the global WorldPop project population dataset, provided at 100 m resolution, allowing for detailed analysis of population density in each region of the world. [12]

- The European Space Agency (ESA) WorldCover 10 m land use dataset for 2021 is cited in GEE under ee.ImageCollection("ESA/WorldCover/v200"). This is a dataset that presents a global land cover map for 2021 at 10 m resolution based on Sentinel-1 and Sentinel-2 data with 11 land cover types. [13]

b. Cloud computing and sample data set establishment

Currently, the analysis and processing of large satellite image data over a long period has become a pressing need in many fields, especially in environmental science [17, 18]. One powerfultool widely used for this purpose is the Google Earth Engine (GEE) cloud computing platformfor analyzing, processing, and computing geospatial data with a metadata repository consistingof raw satellite imagery, ultra-processed imagery, climate maps, climate data, terrain, population data, etc. [19-22]. The GEE cloud computing platform supports various tools, such as statistical tools, image processing tools, and learning machines, supporting the analysis and processing of huge volumes of data in a short time [23, 24].

In order to construct an image interpretation, test sample data must be created using the Stratified Random Sampling (SRS) method based on a certain number of samples selected for each data class. These sampling points will be divided into 2 sample groups: Samples for machine learning training (training) account for 80% of the total number of samples, and control samples (accounting for 20% of the total number of samples) to evaluate the reliability classification results. The study selected random samples throughout the region with a total number of selected - 9,741 samples, of which the number of samples used for machine learning training was 7.740, and the number of control test samples was 2.001. Figure 2 shows the diagram and distribution of random sampling points for the study area:

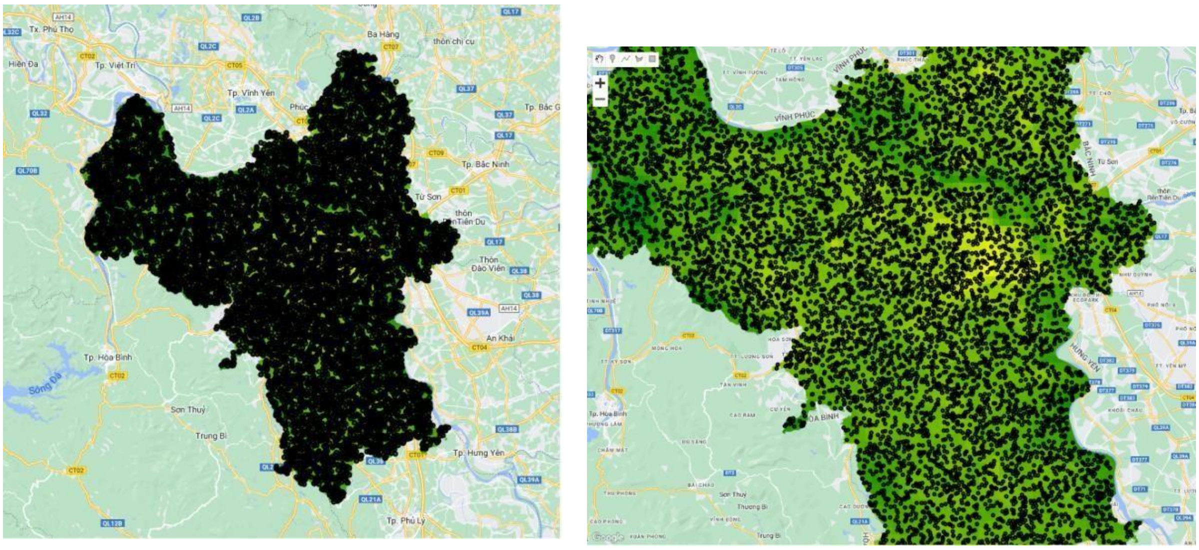

Figure 2. Location and distribution of random sampling points for machine learning

Random Forest algorithm

Random Forest (RF) is a statistical machine learning method proposed by Breiman [1] in a group of combinatorial learning models, namely the multiple decision tree model (decision trees) [2]. Decision trees are a simple way of representing protocols, where each branch of the tree represents attributes and selected values for those attributes. In particular, the decision tree algorithm allows the simultaneous use of classification and regression tasks to specify the predicted value. A random forest classifier is developed based on multiple decision trees created using different random subsets of the data. Each decision tree provides predictions for classification. Random Forest will rely on the majority of prediction results to select the most popular result as the final model output [25, 26].

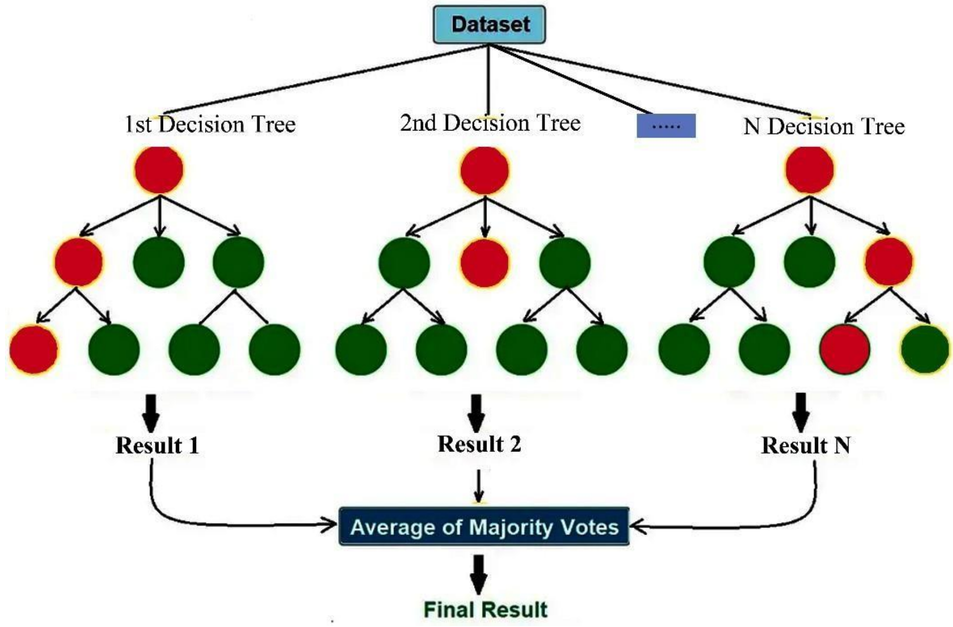

Figure 3. The Random Forest (RF) algorithm

3. Results and discussion

Figure 4 shows the distribution of population density and classification of land cover types in the Hanoi city area. It can be seen that Hanoi is one of the most densely populated urban areas and is still experiencing rapid growth. However, the population is unevenly distributed, resulting in differences in density between areas and between urban and rural areas. The population is concentrated mainly in the urban districts of Hanoi, most of which are urban districts such as Dong Da, Thanh Xuan, Ba Dinh, Hoan Kiem, and Hai Ba Trung, with a density of 210 people/ha; meanwhile, suburban areas such as Ba Vi and My Duc districts have low population densities, less than 10 people/ha. The significant difference in population density between the two areas shows that most of the population is concentrated in the city center with many amenities and employment opportunities. High population densities also increase the need for housing and infrastructure, so there is a transition between land cover types with a gradual loss of vegetation cover and an increase in the built-up area. The area of each land cover type is calculated directly on the cloud platform and exported to Excel files, as shown in Table 1.

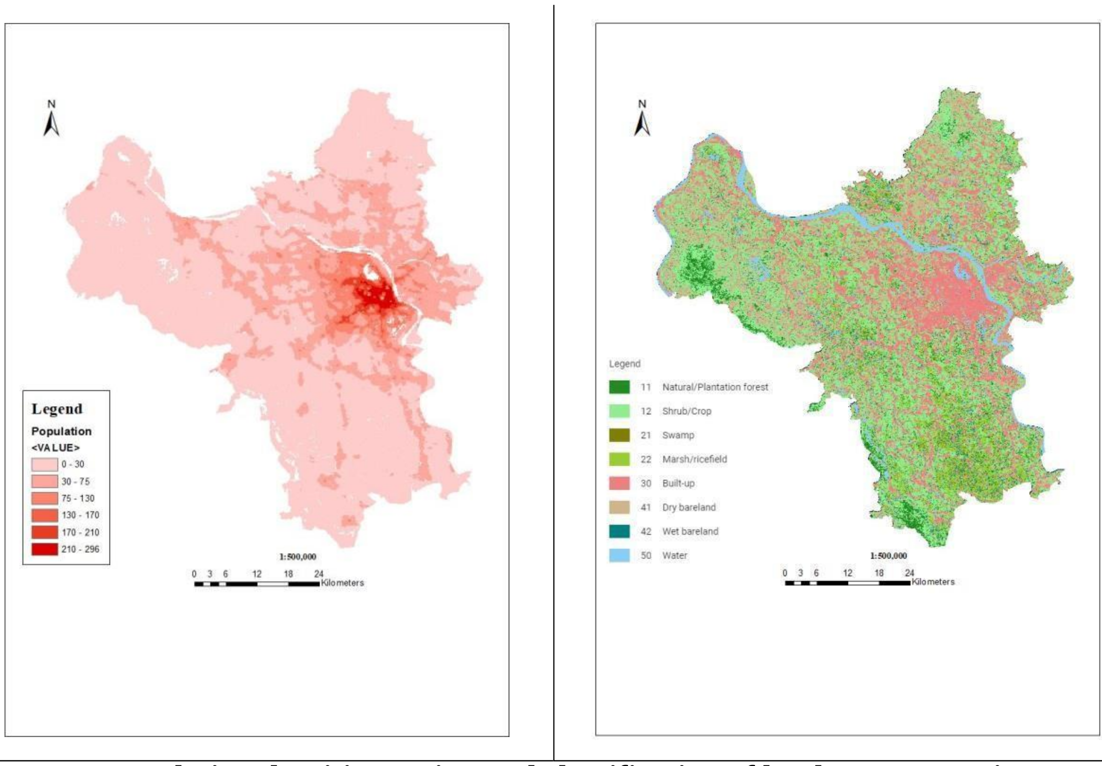

Figure 4. Population densities zoning and classification of land cover types in Hanoi city area

Table 1. Detailed classification of land cover areas in Hanoi

№	**Land cover**	**Area**	**Unit**
1	Tree cover	112.938	Ha
2	Shrubland	45	Ha
3	Grassland	14.669	Ha
4	Cropland	123.245	Ha
5	Built-up	61.715	Ha
6	Bare / sparse vegetation	5.548	Ha
7	Permanent water bodies	17.089	Ha
8	Herbaceous wetland	210	Ha

The results of the random forest model evaluating the influence and importanceof 4 input data factors indicate the following: population density have the mostsignificant impact on LST variable values, followed by precipitation, elevation, and finally, land cover types. Figure 5 illustrates the surface temperature zoning of the studyarea, allowing for a more detailed assessment while considering other influencing factors. Figure 6 shows the relationship between predicted LST and actual LST values.There is a moderate positive correlation between actual and predicted LST values, as indicated by the clustering of points around the regression line. The analysis indicates that the predictive model performs moderately well, capturing the general trend between actual and predicted LST values.

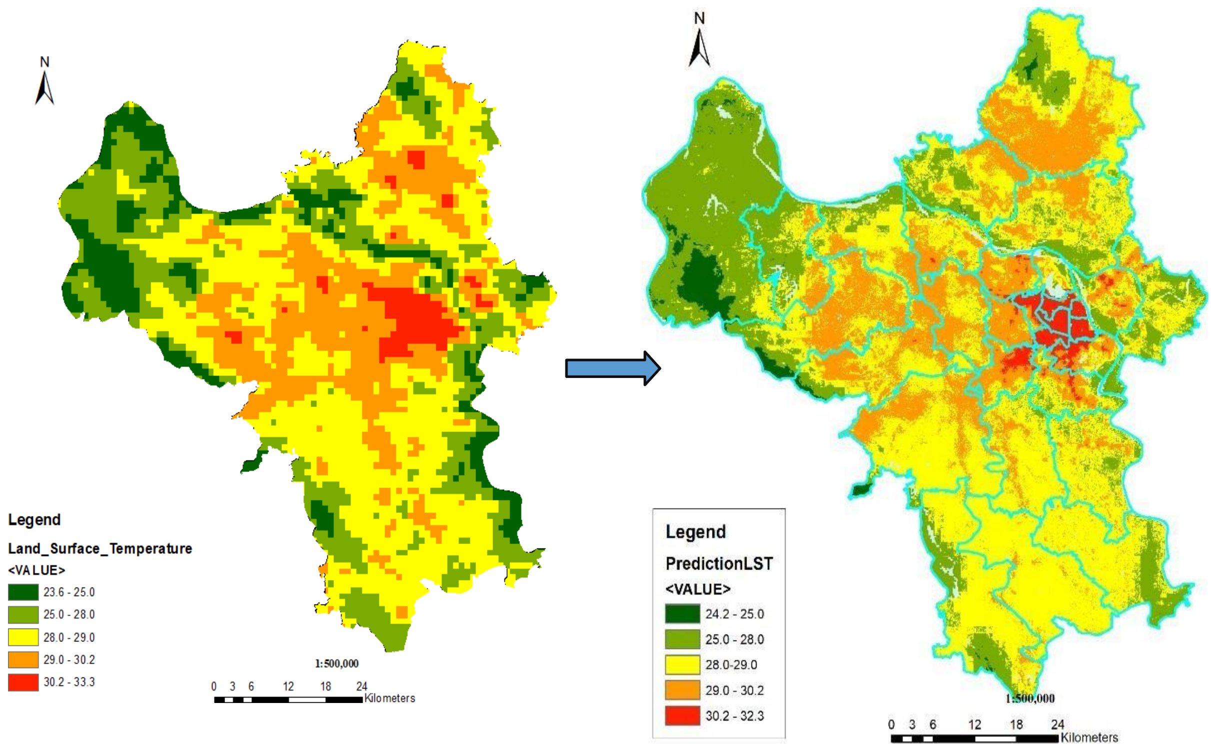

Figure 5. Land surface temperature in the Hanoi city: A) LST according to the data set and B) LST predicted after applying RF model

Table 2. Sampling parameters and correlation values

Parameter	Value
Number of samples	9.741
Training sample	7.740
Comparison inspection sample	2.001
R-square	0,66
Mean LST prediction	28,345 °C

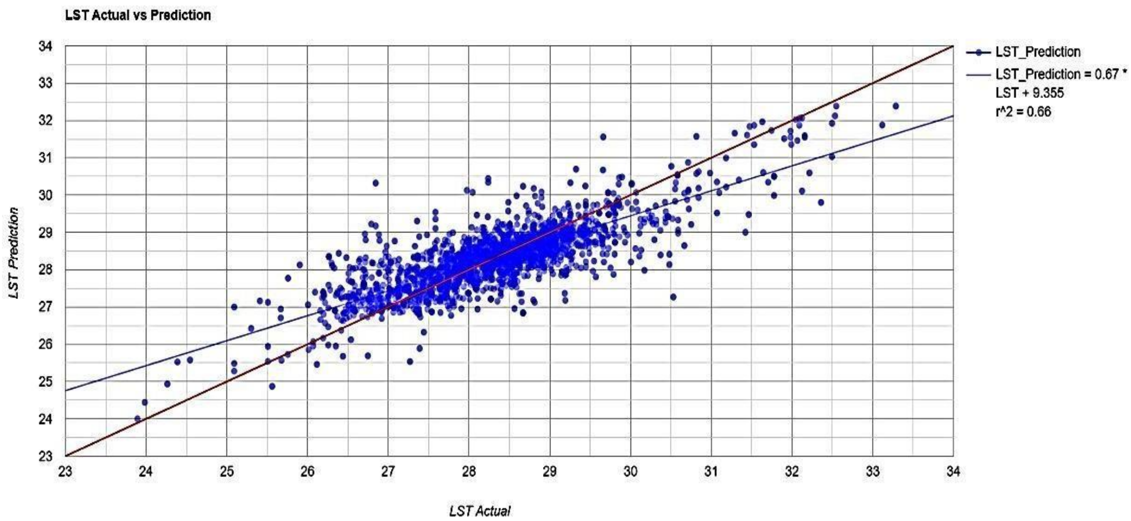

Figure 6. Correlation between LST prediction values and actual values

To evaluate urban heat island zoning, the study used the normalized urban heat island index (Normalized UHI) calculated according to the following formula:

$$UHI_N = \frac{LST - LSTmean}{SD_{LST}} \qquad (1)$$

In which: UHI_N – Standardized Urban Heat Island Index;
LST – Prediction land surface temperature value;
LST_{mean} – Average prediction land surface temperature value;SD_{LST} – Standard deviation value.

The results of urban heat island zoning in Hanoi city in Figure 7 show that the city center area, especially in districts such as Dong Da, Hoan Kiem, Hai Ba Trung, Ba Dinh, and Cau Giay, where urban heat island occurrence is the highest with UHI_N index 2 – 4 °C above the average surface temperature throughout the entire territory.

Areas near the center, such as Bac Tu Liem, Nam Tu Liem, Ha Dong, Thanh Tri, and Tay Ho districts, also show a higher increase in regional temperature than other rural and suburban areas from 1 - 2 °C.

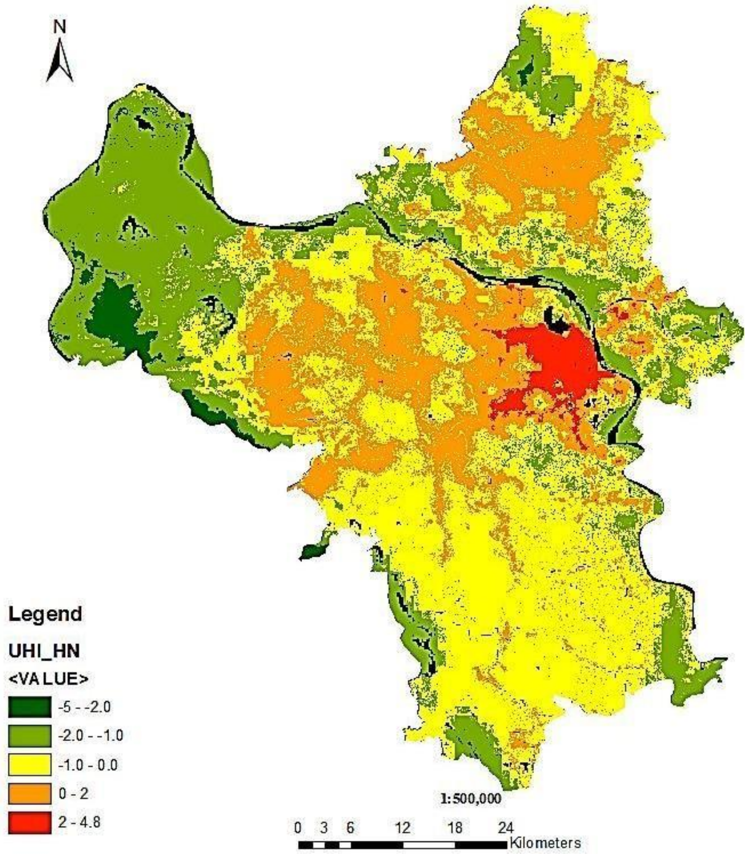

Figure 7. Forecast of urban heat island zoning in the Hanoi city area

4. Conclusion

The prediction of LST results, when compared with LST using the MODIS dataset, achieves a confidence level of R^2 = 0.66 (66 %), which is well suited to the requirements of database generation and an ambient temperature map. Thus, using cloud computing and random forest machine learning models will help to obtain a more detailed land surface temperature dataset using various input factors such as land surface temperature datasets, population, rainfall, and elevation data sets.

The application of machine learning model and cloud computing to predict the urban heat island zoning in Hanoi city not only clarifies the area of influence of the urban heat island phenomenon but also allows management and planning managers to have a more detailed understanding of the sustainable management of the city. and development. In the future, the combination of modern technology and rich data will be an essential key to improving the quality of life, solving environmental problems, and creating sustainable urban development.

Acknowledgments: The work presented in this paper is based on research support by project: "Research and application of simulation model on the impact of the heat island effect on the atmosphere in urban areas of Hanoi city and proposed mitigation solutions" under contract code B2022-MDA-12. The authors are extremely grateful for this support.

References

1. Breiman L. (2001), "Random forests", Machine learning Volume 45, pp. 5-32.

2. Minh Hai Pham and Ngoc Quang Nguyen (2019). An introduction of Random forest in the machine learning revolution and the application in satellite image classification. Journal of Geodesy and Cartography, (39), 15–19. https://doi.org/10.54491/jgac.2019.39.344.

3. Hung T. L. (2018). Combined use of Landsat 8 and Sentinel 2 images for enhanced spatial resolution of land surface temperature. VNU Journal of Science: Earth and Environmental Sciences. 34(4).

4. Quoc Hung Le, Thi Phuong Thao Vu and Thu Huyen Tran (2019). Ability of carbon emission estimation in the field of land use, landuse change and forestry by using remote sensing data. Journal of Geodesy and Cartography, (42), 44–50. https://doi.org/10.54491/jgac.2019.42.357

5. Nhu Duan Dang, Ngoc Long Dao, Le Hung Trinh (2017). Study on the change of land surface temperature in Thanh Hoa city in the period of 2000 – 2017 using Landsat thermal infrared data. Journal of Geodesy and Cartography, (34), 26–32. https://doi.org/10.54491/jgac.2017.34.255.

6. Hung Le Trinh (2014). Study the surface temperature distribution using LANDSAT multispectral satellite image data. Journal of Earth Sciences, Volume 36, Number 01, pages 82 - 89.

7. Thanh Quang Bui (2015). Urban heat island analysis in Ha Noi: examining the relatioship between land surface temperature and impervious surface. National GIS Application Workshop 2015, pages 674 - 677.

8. Nguyễn Đức Thuận, Phạm Văn Vân (2016). Ứng dụng công nghệ viễn thám và hệ thống thông tin địa lý nghiên cứu thay đổi nhiệt độ bề mặt 12 quận nội thành, thành phố Hà Nội giai đoạn 2005 – 2015, Tạp chí Khoa học Nông nghiệp Việt Nam, tập 14, số 8, trang 1219 – 1230.

9. Boori, M.S.; Vozenilek, V.; Balter, H.; Choudhary, K. (2015). Land surface temperature with land cover classes in Aster and Landsat data, Journal of Remote Sensing & GIS 4:138. doi:10.4172/2169-0049.1000138.

10. Cueto, G.; Ostos, J.; Toudert, D.; Martinez, T. (2007). Detection of the urban heat island inMexicali and its relationship with land use, Atmosfera 20(2), pp. 111 – 131.

11. NASA JPL (2020). NASADEM Merged DEM Global 1 arc second V001 [Data set]. NASA EOSDIS Land Processes DAAC. Accessed 2020-12-30 from doi:10.5067/MEaSUREs/NASADEM/NASADEM_HGT.001

12. Asia population count data: Gaughan, A.E.; Stevens, F.R.; Linard, C.; Jia, P. and Tatem, A.J. (2013). High resolution population distribution maps for Southeast Asia in 2010 and 2015, PLoS ONE, 8(2): e55882.

13. Zanaga, D.; Van De Kerchove, R.; Daems, D.; De Keersmaecker, W.; Brockmann, C.; Kirches, G.; Wevers, J.; Cartus, O.; Santoro, M.; Fritz, S.; Lesiv, M.; Herold, M.; Tsendbazar, N.E.; Xu, P.; Ramoino, F.; Arino, O. (2022). ESA WorldCover 10 m 2021 v200. (doi:10.5281/zenodo.7254221)

14. Le, M.T.; Cao, T.A.T.; Tran, N.A.Q (2019). The role of green space in the urbanization of Hanoi city. E3S Web of conferences. - 2019. - Vol. 97. 01013.

15. Fu, P.; Weng, Q. (2016). A time series analysis of urbanization induced land use and land cover change and its impact on land surface temperature with Landsat imagery. Remote Sensing of Environment 175, 205-214.

16. Nguyễn Phương Đông, Phan Thị Mai Hoa, Nguyễn Thị Hòa (2024). Trực quan hóa diễn biến nhiệt độ bề mặt và đảo nhiệt đô thị của thành phố Hà Nội bằng Google Earth Engine và nền tảng điện toán đám mây. Tạp chí Rừng và Môi trường, ISSN: 1859-1248, Vol. 121, pp. 104-107.

17. Amani, M.; Ghorbanian, A.; Ahmadi, S.A.; Kakooei, M.; Moghimi, A.; Mirmazloumi, S.M.; Moghaddam, S.H.A.; Mahdavi, S.; Ghahremanloo, M.; Parsian, S.; et al. (2020). Google Earth Engine cloud computing platform for remote sensing big data applications: A comprehensive review. IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens. 2020, 13, 5326–5350.

18. Xiong, J.; Thenkabail, P.S.; Tilton, J.C.; Gumma, M.K.; Teluguntla, P.; Oliphant, A.; Congalton, R.G.; Yadav, K.; Gorelick, N. (2017). Nominal 30-m cropland extent map of continental Africa by integrating pixel-based and object-based algorithms using Sentinel-2 and Landsat-8 data on Google Earth Engine. Remote Sens. 2017, 9, 1065.

19. Shelestov, A.; Lavreniuk, M.; Kussul, N.; Novikov, A.; Skakun, S. (2017). Exploring Google Earth Engine platform for big data processing: lassification of multi-temporal satellite imagery for crop mapping. Front. Earth Sci. 2017, 5, 1–10.

20. Huang, H.; Chen, Y.; Clinton, N.; Wang, J.; Wang, X.; Liu, C.; Gong, P.; Yang, J.; Bai, Y.; Zheng, Y.; et al. (2017). Mapping major land cover dynamics in Beijing using all Landsat images in Google Earth Engine. Remote Sens. Environ. 2017, 202, 166–176.

21. Le, M. T.; Bakaeva N. (2023). A Technique for Generating Preliminary Satellite Data to Evaluate SUHI Using Cloud Computing: A Case Study in Moscow, Russia", Remote Sensing 2023, 15(13): 3294.

https://doi.org/10.3390/rs15133294

22. Hao, B.; Ma, M.; Li, S.; Li, Q.; Hao, D.; Huang, J.; Ge, Z.; Yang, H.; Han, X. Land use change and climate variation in the three gorges reservoir catchment from 2000 to 2015 based on the Google Earth Engine. Sensors 2019, 19, 2118.

23. Tamiminia, H.; Salehi, B.; Mahdianpari, M.; Quackenbush, L.; Adeli, S.; Brisco, B. (2020). Google Earth Engine for geo-big data applications: A meta-analysis and systematic review. ISPRS journal of photogrammetry and remote sensing, 164, 152-170.

24. Agbehadji, I. E.; Mabhaudhi, T.; Botai, J.; Masinde, M. (2023). A systematic review of existing early warning systems' challenges and opportunities in cloud computing early warning systems. Climate 2023, 11(9), 188.

https://doi.org/10.3390/cli11090188

25. Teluguntla, P.; Thenkabail, P.S.; Oliphant, A.; Xiong, J.; Gumma, M.K.; Congalton, R.G.; Yadav, K.; Huete, A. A 30-m landsat-derived cropland extent product of Australia and China using random forest machine learning algorithm on Google Earth Engine cloud computing platform. ISPRS-J. Photogramm. Remote Sens. 2018, 144, 325–340.

26. Hengl, T.; Nussbaum, M.; Wright, M. N.; Heuvelink, G. B.; Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ, 6, e5518.