

Jun 19th, 10:50 AM - 11:50 AM

Evaluating the Evaluation Matrices: Integrating Spatial Assessment in Geospatial AI Model Training and Evaluation

Fangzheng Lyu

Virginia Polytechnic Institute and State University, fangzheng@vt.edu

Follow this and additional works at: <https://docs.lib.purdue.edu/iguide>



Part of the [Artificial Intelligence and Robotics Commons](#), and the [Geographic Information Sciences Commons](#)

Recommended Citation

Lyu, Fangzheng, "Evaluating the Evaluation Matrices: Integrating Spatial Assessment in Geospatial AI Model Training and Evaluation" (2025). *I-GUIDE Forum*. 5.
<https://docs.lib.purdue.edu/iguide/2025/presentations/5>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries.
Please contact epubs@purdue.edu for additional information.

Evaluating the Evaluation Matrices: Integrating Spatial Assessment in Geospatial AI Model Training and Evaluation

Fangzheng Lyu
Department of Geography
Virginia Tech
Blacksburg, VA, United States
fangzheng@vt.edu

Abstract— This paper examines the limitations of current evaluation metrics in GeoAI. Through two case studies on deep learning models—a building detection classification problem and a remote sensing image fusion regression problem—this paper demonstrates how traditional statistical evaluation matrices alone can be misleading in geospatial problems. The findings indicate that traditional metrics (e.g., RMSE, MAE) used in current GeoAI models can have difficulty capturing the spatial dimensions inherent to geospatial problems. This paper suggests that the model evaluation process in GeoAI should move beyond traditional evaluation matrices by integrating spatial thinking throughout the modeling pipeline—not only incorporating spatial accuracy in model evaluation but also embedding it within optimization functions in model structure and model training.

Keywords—*Geospatial AI, Evaluation Matrices, GIS*

I. INTRODUCTION

Geospatial AI extends beyond merely applying AI to geospatial problems/data; it represents the integration of spatial thinking with modern AI tools. However, the geospatial models, particularly those related to Geospatial Artificial Intelligence (GeoAI), mostly utilize traditional statistical evaluation metrics (e.g., Root Mean Square Error (RMSE), Mean Absolute Error (MAE)) exclusively for model training and model performance evaluation. Widely adopted by the computer science community, deep learning classification problems commonly used evaluation metrics include classification accuracy, F1 score, and Area Under Curve (AUC). For deep learning regression problems, evaluation metrics such as MAE and RMSE are frequently employed. As the computer science community leads machine learning and AI model development, recent deep learning models—including transformer-based language models, vision transformers, diffusion models, and ResNet—when adopted by the geospatial community, utilize deep learning structures and evaluation metrics developed by computer scientists and apply them to geospatial problems and data. Admittedly, these commonly adopted statistical evaluation metrics provide a standard and well-accepted method (recognized by both the geospatial community and the broader scientific community) for assessing and evaluating geospatial model performance. However, geospatial problems are unique in that spatial accuracy cannot be fully captured by traditional evaluation metrics that focus primarily on object classification accuracy or regression error. Geospatial problems and models

require more than pixel-wise accuracy achieved by minimizing statistical evaluation metrics such as RMSE. The overall spatial accuracy plays a crucial role in assessing model performance, especially for the AI-based model within a geospatial context. In this paper, the author suggests that while statistical evaluation metrics remain significant in geospatial model assessment, additional effort needed be devoted to evaluating spatial accuracy to ensure it is adequately addressed and that model assessment incorporates spatial context. Spatial accuracy should not only be considered in the model assessment and evaluation stage, but also could be incorporated in the objective function in the model training stage. In the following sections, the author will use two case examples representing typical deep learning classification and regression problems in GeoAI to evaluate the evaluation metrics.

II. CLASSIFICATION PROBLEM

We use building detection as an example for a classification problem to look at common evaluation matrices in geospatial deep learning model. In this case study, we aim to use deep learning models and street view data to detect new building construction between 2007 and 2023 using Google Street View data. We collected 102,112 images, containing 14,272 sets of time-series street view images from 3,577 locations distributed across Mecklenburg County, North Carolina, where the city of Charlotte is located [1], [2]. The principle for site selection is based on the census block group in 2010, and 10 sites within the census block is randomly selected and keeping all the valid sites with Google Street View images. And among these locations, we identified 558 sites out of 3,577 sites with building construction. Leveraging geospatial AI, we developed deep learning models based on Vision Transformer architecture and Convolution Neural Network (CNN) to categorize street view images captured at different locations and time periods to detect building construction in Mecklenburg County, North Carolina. As a typical deep learning classification problem with classification accuracy as the evaluation matrices, this task also has spatial dimensions, where both the values in the confusion matrix as well as the spatial distribution of new building construction matter. Our analysis generated two deep learning models – one with CNN and the second with Vision Transformer - with identical confusion matrices (see Table 1). And for both models, using traditional statistical evaluation

matrix, the prediction accuracy (0.9956) and F1-score (0.9839) are the same.

TABLE I. MODEL CONFUSION MATRIX

	Actually Positive	Actually Negative
Predicted Positive	550	10
Predicted Negative	8	3678

Based on the evaluation for statistical accuracy, the performance of these two models should be identical based on the confusion matrix (see Table 1). However, assessing the models from a spatial perspective yields different perspectives. Figure 1a) shows the distribution of new construction identified in the street view data as ground truth for reference. Compared with the periphery of Mecklenburg County, new construction clusters near downtown Charlotte. Figure 1b) and Figure 1c) show the spatial distribution of misclassification predicted by two different models developed for this project – false positive indicating false new construction predicted by deep learning model while false negative representing falsely classified no new construction predicted by the deep learning model. Examining the distribution of misclassification reveals that the first deep learning model produces misclassifications across the entire county, with false positive mostly distributed in the southern part of the county. Meanwhile, the second deep learning model has misclassifications primarily near the downtown area, indicating less optimal result predicting building construction in that region. Despite the two models generating identical evaluation metrics in statistical matrices based on the confusion matrices (see Table 1), the spatial accuracy of their predictions differs. For example, when analyzing spatial autocorrelation using global Moran's I to examine the spatial clustering of new building construction, we find that the reference data exhibits a Moran's I value of 0.146, while the first model generates a Moran's I of 0.075 and the second model #2 results in a Moran's I of 0.071. This indicates that when evaluating spatial autocorrelation of model performance, the first model demonstrates better performance as its Moran's I value more closely approximates that of the reference data – showing better spatial accuracy. By visual inspection, the distribution of misclassification in the second model indicates model incompatibility for new construction detection in data located near downtown, suggesting that a potential different deep learning model should be adopted specifically for the downtown regions. The first law of geography indicates that near things are more related than distant things, but sometimes physical distance is not an optimal measurement for proximity—in this case, urban planning and zoning differences between downtown area and non-downtown area lead to distinct patterns of new construction between these two areas, despite their relative proximity in Euclidean distance. For example, in this project, developing one deep learning model might not be an optimal solution. A strategy to implement a region separation to delineate the two areas—applying the first deep learning model to downtown regions and the second deep learning model to other parts of the county—would potentially enhance both the spatial accuracy and classification accuracy of the models. This spatial-awareness approach to deep learning model evaluation demonstrates that acknowledging underlying

geographic contexts can improve overall model performance beyond what traditional statistical metrics can offer.

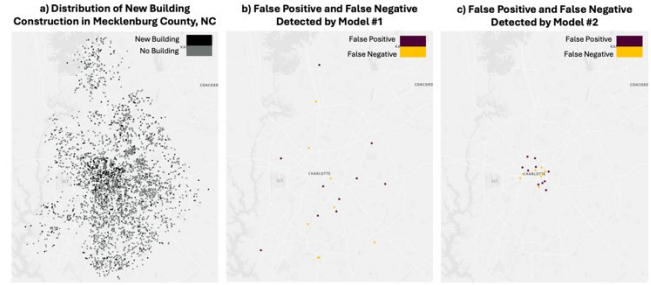


Fig. 1. Distribution of new building construction and misclassification using two deep learning model.

In classification problems, classification accuracy and the confusion matrix typically serve as the model evaluation tools. However, the spatial distribution of correct classifications and misclassifications also matter in classification problems within a spatial context. This could be achieved by not only incorporating accuracy and spatial assessment indicators when evaluating deep learning model performance but also including spatial indicators in the model training process as evaluation metrics to be optimized.

III. REGRESSION PROBLEM

Remote sensing image fusion aims to generate high-spatial-high-temporal resolution remote sensing images based on existing high-spatial-low-temporal and low-spatial-high-temporal resolution remote sensing data [3]. A typical application involves generating high-spatial-high-temporal resolution remote sensing data with high-spatial-low-temporal Landsat imagery and low-spatial-high-temporal resolution MODIS imagery. This process of remote sensing image fusion frequently involves deep learning models and represents a typical regression-based deep learning application in geospatial problems.

The evaluation and assessment of remote sensing image fusion models requires multiple levels of assessment, not only examining spectral accuracy (pixel-wise evaluation) but also considering spatial perspectives [4]. We utilize the benchmark data at an agricultural production area - Coleambally Irrigation Area (CIA) data located in southern New South Wales, Australia [5] - to test the performance of four benchmark deep learning models for remote sensing image fusion: 1) Flexible Spatiotemporal Data Fusion (FSDAF) [6]; 2) GAN-based Spatiotemporal Fusion Model (GAN-STFM) [7]; 3) High-Precision Remote Sensing Spatiotemporal Fusion Method (HPLTS-GAN) [8]; and 4) Enhanced deep convolutional spatiotemporal fusion network (EDCFTSN) [9]. Figure 2 presents the prediction results for April 11, 2004. From the perspective of spectral accuracy, we found that HPLTS-GAN and GAN-STFM performed better in the zoomed detail areas, indicating higher spectral accuracy, while FSDAF and EDCFTSN demonstrated higher error compared to the reference data [10]. Meanwhile, from the perspective of spatial accuracy, EDCFTSN exhibited less optimal performance in the zoomed details as the model had difficulties predicting the red (agriculture) areas.

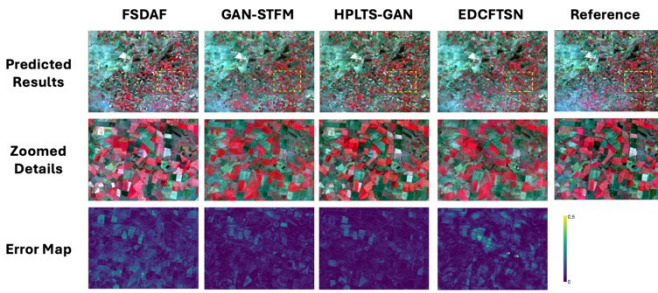


Fig. 2. Remote sensing image fusion model result.

Looking at the evaluation metrics using spectral accuracy and spatial accuracy, we employed multiple measures for comprehensive assessment (see Table 2). Regarding spectral accuracy, evaluation metrics including RMSE, relative dimensionless global error synthesis (ERGAS), spectral angle mapping (SAM), and structural similarity index measure (SSIM) were used to determine model performance from the spectral perspective [11]. On the other hand, to assess spatial accuracy, local binary patterns (LBP) metrics were utilized [12]. Table 2 presents the results of spectral and spatial accuracy using these five evaluation metrics. Based on the model comparisons, if we evaluate the model performance based on traditional metrics such as RMSE or general spectral evaluation metrics, we would conclude that EDCSTFN and FSDAF demonstrate the best performance. However, when incorporating spatial accuracy into consideration, we find that FSDAF does not generate satisfactory results for spatial accuracy. And EDCSTFN is believed to be the best-performing model, effectively balancing both spatial and spectral accuracy requirements.

TABLE II. SPECTRAL ACCURACY AND SPATIAL ACCURACY

	Spectral				Spatial
	RMSE	ERGAS	SAM	SSIM	LBP
FSDAF	0.0331	1.3001	0.0995	0.9390	0.1259
GANSTFM	0.0504	1.9117	0.1327	0.8524	0.1235
HPLTS-GAN	0.0420	1.3100	0.1220	0.9250	0.1220
EDCSTFN	0.0371	1.2896	0.0897	0.9357	0.1233

In this typical regression-based deep learning model, not only is spectral (pixel-wise) accuracy significant, but spatial accuracy is also crucial in model evaluation. However, existing benchmark models, despite being evaluated on spatial accuracy, do not incorporate spatial accuracy in the optimization function of the deep learning model training process. This leads to a situation where, despite the concept of comprehensive evaluation metrics being present, the model structure and model training process still ignore spatial accuracy.

IV. CONCLUSION & FUTURE WORK

Geospatial AI faces an identity crisis in establishing itself as a unique discipline that contributes to the scientific community beyond applying AI algorithms to geospatial data. A significant limitation exists in the evaluation metrics and systems

employed—current geospatial AI models mostly rely on traditional statistical evaluation metrics such as RMSE that are widely adopted by the computer science community but lack evaluation from a spatial perspective. This paper evaluates the current state of evaluation metrics used in Geospatial AI through two examples of deep learning classification and regression problems, suggesting that geospatial analysis should inherently incorporate spatial dimensions and assess spatial accuracy in evaluation. This approach extends beyond simply introducing spatial accuracy metrics in the evaluation matrices (e.g., LBP, SSIM) when assessing model performance, suggesting for the integration of spatial accuracy within the model structure and model training process for model optimization. Future work could focus on new evaluation metrics development or optimization techniques by the geography community to better address the specific requirements of geospatial AI problems from the spatial aspect. Additionally, the concept of distance, extending beyond Euclidean distance, could be involved in spatial problems and spatial accuracy assessment, particularly for integration into geospatial AI models—representing a promising direction for future research.

REFERENCES

- [1] F. Lyu, X. Ma, Y. Song, E. Zhu, and S. Wang, "Large-scale Google Street View Images for Urban Change Detection," in *I-GUIDE Forum 2023*, 2023.
- [2] X. Ma, Y. Song, F. Lyu, Y. Yang, X. Li, and S. Zhong, "Revitalizing Cities: The SR Framework Approach to Urban Retrofitting and Big Data Insights." *Growth and Change*, 56(1), p.e70018, 2025.
- [3] F. Lyu, Z. Yang, Z. Xiao, C. Diao, J. Park, and S. Wang, "CyberGIS for Scalable Remote Sensing Data Fusion," in *Practice and Experience in Advanced Research Computing*, Boston MA USA, 2022.
- [4] X. Zhu *et al.*, "A novel framework to assess all-round performances of spatiotemporal fusion models," *Remote Sens. Environ.*, vol. 274, p. 113002, Jun. 2022.
- [5] I. V. Emelyanova, T. R. McVicar, T. G. Van Niel, L. T. Li, and A. I. J. M. van Dijk, "Assessing the accuracy of blending Landsat–MODIS surface reflectances in two landscapes with contrasting spatial and temporal dynamics: A framework for algorithm selection," *Remote Sens. Environ.*, vol. 133, pp. 193–209, Jun. 2013.
- [6] X. Zhu, E. H. Helmer, F. Gao, D. Liu, J. Chen, and M. A. Lefsky, "A flexible spatiotemporal method for fusing satellite images with different resolutions," *Remote Sens. Environ.*, vol. 172, pp. 165–177, Jan. 2016.
- [7] Z. Tan, M. Gao, X. Li, and L. Jiang, "A Flexible Reference-Insensitive Spatiotemporal Fusion Model for Remote Sensing Images Using Conditional Generative Adversarial Network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.
- [8] D. Lei *et al.*, "HPLTS-GAN: A high-precision remote sensing spatiotemporal fusion method based on low temporal sensitivity," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–16, 2024.
- [9] Z. Tan, L. Di, M. Zhang, L. Guo, and M. Gao, "An enhanced deep convolutional model for spatiotemporal image fusion," *Remote Sens. (Basel)*, vol. 11, no. 24, p. 2898, Dec. 2019.
- [10] F. Lyu, Z. Yang, C. Diao, and S. Wang, "Multistream STGAN: A Spatiotemporal Image Fusion Model With Improved Temporal Transferability," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 18, pp. 1562–1576, 2025.
- [11] L. Wald, "Quality of high resolution synthesised images: Is there a simple criterion?," in *Third conference" Fusion of Earth data: merging point measurements, raster maps and remotely sensed images"*, 2000, pp. 99–103.
- [12] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.